CENTRO INTERNACIONAL DE ESTUDOS
DE DOUTORAMENTO E AVANZADOS
DA USC (CIEDUS)

TESE DE DOUTORAMENTO

CONTRIBUTIONS TO DISTRIBUTIONAL REGRESSION MODELS. APPLICATIONS IN BIOMEDICINE

Jenifer Espasandín Domínguez

ESCOLA DE DOUTORAMENTO INTERNACIONAL

PROGRAMA DE DOUTORAMENTO EN ESTATÍSTICA E INVESTIGACIÓN OPERATIVA

SANTIAGO DE COMPOSTELA

2019

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

DECLARACIÓN DA AUTORA DA TESE

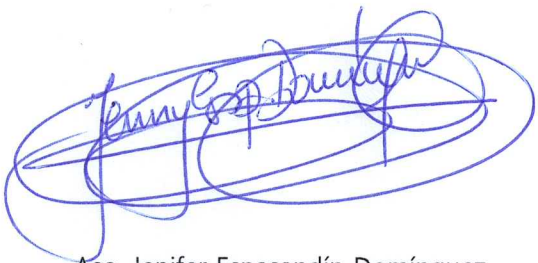
"Contributions to distributional regression models. Applications in biomedicine"

D./Dna. JENIFER ESPASANDÍN DOMÍNGUEZ

Presento a miña tese, seguindo o procedemento axeitado ao Regulamento, e declaro que:

1) A tese abarca os resultados da elaboración do meu traballo.
2) De selo caso, na tese faise referencia ás colaboracións que tivo este traballo.
3) A tese é a versión definitiva presentada para a súa defensa e coincide coa versión enviada en formato electrónico.
4) Confirmo que a tese non incorre en ningún tipo de plaxio doutros autores nin de traballos presentados por min para a obtención doutros títulos.

En Santiago de Compostela, 03 de Maio de 2019



Aso. Jenifer Espasandín Domínguez

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

AUTORIZACIÓN DO DIRECTOR / TITOR DA TESE

CONTRIBUTIONS TO DISTRIBUTIONAL REGRESSION MODELS. APPLICATIONS IN BIOMEDICINE

Dna. Carmen María Cadarso Suárez
D. Francisco Gude Sampedro
D. Thomas Kneib

INFORMAN:

Que a presente tese, correspóndese co traballo realizado por Dna. ***Jenifer Espasandín Domínguez****, baixo a nosa dirección, e autorizamos a súa presentación, considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC, e que como directores desta non incorre nas causas de abstención establecidas na Lei 40/2015.*

En Santiago de Compostela, 5 de Maio de 2019.

Asdo. Carmen María Cadarso Suárez

Asdo. Francisco Gude Sampedro

Asdo Thomas Kneib

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Universidade de Santiago de Compostela

DISSERTATION

CONTRIBUTIONS TO DISTRIBUTIONAL REGRESSION MODELS. APPLICATIONS IN BIOMEDICINE

Author:
Jenifer Espasandín Domínguez

Advisors:
Carmen Cadarso-Suárez
Francisco Gude
Thomas Kneib

DEPARTMENT OF STATISTICS, MATHEMATICAL ANALYSIS, AND OPTIMIZATION

Santiago de Compostela, 2019

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Dedícocha a ti, madriña, sempre seguirás viva comigo

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Agradecementos-Acknowledgments

O desenvolvemento desta tese non foi unha tarefa doada, pero si podo dicir que tiven a sorte de estar rodeada de grandes persoas ás que me gustaría agradecer profundamente todo o seu apoio. Grazas a todos! Este traballo tamén é voso.

Gustaríame comezar amosando o meu profundo agradecemento aos meus directores de tese, Carmen Cadarso, Francisco Gude e Thomas Kneib. Moitas grazas por todo o voso apoio e axuda nesta etapa. Moitas grazas, por ser algo máis que directores académicos. A vosa pegada traspasa ao terreo persoal.

Carmen, grazas por confiar en min dende o primeiro momento e animarme a comezar este traballo de investigación. Grazas por dirixirme e guiarme todos estes anos. Sen ti, ningunha peza do puzzle encaixaría. Por todos os consellos, e ánimos, por entender cada momento complicado e transmitirme a túa paixón investigadora. Por todas as estadías, congresos, cursos e todo o soporte económico. Por motivarme e estar sempre ao meu lado; e por acollerme de forma tan bonita na unidade de Bioestatística, na que coñecín e compartín momentos con persoas marabillosas das que levarei sempre un pedaciño en min.

Francisco, moitas grazas por proporcionarme o material para a realización desta tese doutoral: "os datos", sen os que este proxecto non tería cabida. Moitas grazas pola túa paciencia, polas túas correccións e por todo o tempo adicado a mellorar e dar luz a cada un dos traballos que realicei. E, en definitiva, por acompañarme nesta etapa, foi un pracer para min ser a túa alumna.

Thomas, thank you very much for beeing very closed to me, and for your patient guidance during these years. Thank you very much for your continued support.

I also acknowledge the help of Professors Nadja Klein, Giampiero Marra and Rosalba Radice. This thesis would not be possible without you.

Esta tese, tamén vai adicada, a todos aqueles maestros e profesores que logran inspirar e deixar unha pegada para sempre nos seus alumnos e alumnas. Neste senso, gustaríame facer unha mención especial aos profesores do Instituto Agra de Raíces de Cee, o meu pobo, por transmitirme tanto e animarme sempre. E por suposto, aos meus profesores da Facultade de Matemáticas e do Departamento de Estatística, Análise Matemática e Optimización da USC. E á Xunta de Galicia, polo soporte económico para poder seguir formándome na miña comunidade.

Grazas tamén a todos os meus compañeiros do grupo de Biostatística e Ciencia de Datos Biomédicos da USC: Aos profesores Carmen Carollo e Xosé Luis

Otero por acompañarme e apoiarme nos meus primeiros pasos como docente. Grazas por todos os vosos consellos. A Ipek, pola alegría que sempre desprende. A Ana, por traspasar a barreira de compañeira e convertirse nunha gran amiga e un apoio incondicional. Grazas de corazón. Grazas en nome de todos tamén por todos os trámites e papeleos que sempre nos solucionas. Pola túa eficacia e polas túas sempre acertadas palabras. A Berta, polo sorriso e o abrazo co que sempre nos recibe cada mañá. Grazas ademais por todas as infusións e por alimentar o meu estómago co que máis me gusta.

A Óscar, por aguantarme tantas mañás, tardes e festivos. Por estar sempre presente e disposto a axudar. Grazas polos debates, as dudas e todas as risas. Carliña, moitas grazas por todo o teu apoio emocional e computacional! E sobre todo, moitas grazas por todo o teu tempo. Existen poucas persoas tan intelixentes e xenerosas coma ti. Juan, moitas grazas polas túas leccións de inglés!

A todos os compañeiros e compañeiras de congresos, por amenizar esta etapa e escoitar cada problema, e en definitiva, por facer de cada evento, un recordo especial. A todos os meus amigos e amigas, e en especial, a Anaïs e a Sonia por estar sempre presentes. Grazas especiais tamén para o resto dos meus amigos da *"Residencia Universitaria Monte da Condesa"* e de *"Carreira do Conde"* e por suposto, a Ángeles, vivir comigo esta última tempada, non puido ser (nin é) fácil. E as miñas *"pitagóricas"* preferidas, Iria, María e Noe, o meu paso polo Grao en Matemáticas, non sería o mesmo sen vós.

A Manuel, por acompañarme durante todos estes anos e nunca soltarme a man. Por apoiarme e compartir comigo durante todo este proceso alegrías, e penas, aínda non entendéndoas moitas veces. Quérote moito. Moitas grazas.

A miña familia. Grazas papá e MAMÁ, por todos os vosos esforzos para que eu sexa a persoa que hoxe en día son. Por animarme dende pequena a estudar. Por apoiarme en todos os momentos complicados. Por obrigarme a levantarme despois de cada tropezo e por non permitir que nunca me rendera. Quérovos moito.

Xa para rematar, gustaríame recordar aos meus catro avós. A pesar de que ningún deles puido ver este traballo finalizado, gran parte do mesmo lles pertence. Grazas en especial á miña avoa, tamén madriña, gran confidente e un apoio incondicional, non só nesta etapa predoutoral senón durante toda a miña vida. Grazas por todas as túas caricias, por todo o teu amor, polos teus valiosos consellos e pola túa entrega incondicional. Sinto moito que non poidas ver este libro rematado. Póñoche moito de menos.

Grazas por todo.

Jenifer Espasandín Domínguez
San Xosé de Pereiriña, Cee, Maio 2019

Fundings: This thesis forms part of the *"New healthy lifestyle model"* challenge, focusing on the areas of *"active ageing of population"* and *"Prevention, diagnosis and treatment of diseases"*. This challenge is financed by the programme *"Axudas de apoio á etapa predoutoral do Plan Galego de Investigación, Innovación e Crecemento 2011-2015 (Plan I2C)"*, which provided a grant to the present author in 2015. This research was also supported by grants from the *Carlos III Health Institute*, Spain (*PI16/01395; PI16/01404; RD16/0007/0006 and RD16/0017/0018*), and by the projects *MTM2014-52975-C2-1-R* and *MTM2017-83513-R* cofinanced by the *Ministry of Economy and Competitiveness* (SPAIN) and the *European Regional Development Fund* (FEDER). This work was also supported by grants from the Galician Government: *RED INBIOEST* (ED341D-R2016/032), *Grupo de Referencia Competitiva (ED431C 2016/025)* and *Grupo de Potencial Crecimiento (GPC2013/022)*.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Abstract

The origin of this work lies in the convergence of two lines of research: a statistics research line undertaken by the *Group in Biostatistics and Biomedical Data Science, GI-2127,* and a clinical research line followed by the *Research Methods Group, C017;* both groups belong to the *Instituto de Investigación Sanitaria de Santiago de Compostela*. For many years they have coordinated their work to perform interdisciplinary biostatistical research.

The present thesis makes contributions in statistical methodology to aid research into the factors involved in protein glycation. Earlier projects generated a sample of the general adult population for which extensive phenotypic details are available, as well as stored biological samples that can be used to study chronic diseases related to ageing, such as diabetes. The results of week-long continuous monitoring of interstitial glucose concentrations are also available for some members of this population; glucose profiles are therefore available as functional data for each of these individuals. The present work proposes statistical methods for functional data that take into account all the information contained in glucose curves.

This thesis also discusses frequentist and Bayesian models of distributional regression for univariate responses (Rigby and Stasinopoulos, 2005; Klein et al., 2015). Compared to classical regression models based on the linear estimation of the mean of the response variable, these models allow for great flexibility in modelling the response variable and any possible predictor covariates. The incorporation of reference bands into quantile-quantile plots – as a generalization of Augustin et al. (2012) – is one of the major statistical contribution of the present work in the context of distributional regression models. The use of these plots in model selection is validated via a simulation study.

In some practical situations, it is necessary to model multivariate responses. For example, in the framework of this thesis would be of great interest to simultaneously study the glycation of several proteins (such as glycated haemoglobin and fructosamine) and the factors that could influence such glycation. Tackling the problem of multivariate response modelling thus requires new multivariate regression techniques such as copula distributional regression models, introduced by Klein and Kneib (2016b) in Bayesian, and Marra and Radice (2017a) in the frequestist framework. These techniques are compared for the first time via a simulation study and a real biomedical study.

Functional data analysis techniques (Ramsay and Silverman, 2005) are useful for incorporating the above-mentioned glucose profiles into regression models, but this time by means of entering as covariate in distributional regression models (as defined by Klein et al., 2014a). The present work adapts the techniques of Brockhaus et al. (2018) to allow the incorporation of functional data as covariates into univariate distributional regression, and validates the methodology proposed via a simulation study. For the context of multivariate analysis, an extension of the methodology of McLean et al. (2014) is presented that allows functional covariates to be contemplated in distributional regression models based on copulas proposed by Marra and Radice (2017a). To the best of our knowledge, no other copula regression model exists that allows functional covariates to be modelled in a flexible manner. Both extensions are illustrated in the context of continuous glucose monitoring. The code used in the present work is provided as supplementary material to allow the statistical techniques discussed to be reproduced and used by others.

Resumo

O traballo de investigación realizado na presente tese parte da converxencia das liñas de investigación de dous grupos, un de estadística *(Group in Biostatistics and Biomedical Data Science, GI-2127)* e outro de medicina clínica *(Research methods, C017)*, ambos pertencentes ao Instituto de Investigación Sanitaria de Santiago de Compostela e coordinados dende fai anos para levar a cabo investigación interdisciplinar no ámbito da bioestatística.

Nesta tese, realízanse contribucións metodolóxicas no ámbito da estatística, para a investigación de factores determinantes na glicación de proteínas. Especificamente, como resultado de varios proxectos previos, disponse dunha ampla mostra da poboación xeral adulta, cunha extensa fenotipación e almacenamento de mostras biolóxicas que permite investigar diversos retos actuais no campo das enfermidades crónicas relacionadas co envellecemento da poboación, como a diabetes. Ademais, contouse cos resultados da monitorización continua da glucosa intersticial durante unha semana dunha parte da mostra. Disponse, polo tanto, dos perfiles de glucosa de cada individuo como dato funcional. Nesta tese, presentáronse métodos estatísticos para datos funcionais que permiten ter en conta toda a información contida nas curvas de glucosa.

Especificamente, dende un punto de vista estatístico, nesta investigación revisáronse modelos de regresión distribucional frecuentista e Baiesianos para respostas univariantes (Rigby and Stasinopoulos, 2005; Klein et al., 2015). Estas metodoloxías permiten gran flexibilidade tanto no modelado da variable resposta como nas posibles covariables predictoras, fronte aos modelos de regresión clásica que se basean unicamente na estimación lineal da media da variable resposta. Neste ámbito, na presente tese, incorporáronse bandas de referencia aos quantile-quantile plots, no contexto da regresión distribucional - como xeralización da proposta de Augustin et al. (2012). A adecuación destes gráficos no ámbito da regresión distribucional comprobouse mediante un estudo de simulación.

Nalgúns casos prácticos, é necesario ter en conta máis dunha variable resposta. Por exemplo, no eido desta tese é de especial interese estudar de forma simultánea o comportamento de dúas proteínas glicadas (a fructosamina e a hemoglobina glicada) e os factores que poden influenciar esta glicación. Nesta tese, revisáronse novas técnicas estatísticas no ámbito da regresión multivariante, como os modelos de regresión distribucional de cópula, introducidos no eido

Baiesiano por Klein and Kneib (2016b) e no frecuentista por Marra and Radice (2017a). Ademais, por primeira vez - segundo o noso coñecemento - comparáronse mediante un estudo de simulación e nunha base de datos real.

As ferramentas para a análise de datos funcionais (Ramsay and Dalzell, 1991) son de gran utilidade para poder incorporar nos modelos de regresión, toda a información dispoñible nos perfís de glucosa mencionados. Na presente tese doutoral, adaptáronse as técnicas introducidas por Brockhaus et al. (2018) para permitir a incorporación de datos funcionais como covariables no ámbito da regresión distribucional univariante. Ademais, esta metodoloxía foi validada mediante un estudo de simulación. No contexto multivariante, estendeuse a metodoloxía de McLean et al. (2014) para posibilitar a consideración de covariables funcionais nos modelos de regresión distribucional baseados en cópulas propostos por Marra and Radice (2017a). Segundo o noso coñecemento, ata a data, non existe ningún outro modelo de regresión de cópula que permita modelar covariables funcionais cunha flexibilidade similar. Ambas extensións, empregáronse no contexto da monitorización continua da glucosa. Como material suplementario, proporciónase o código de programación deseñado durante o transcurso da presente tese doutoral.

Contents

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

List of Figures

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

List of Tables

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 1

Introduction

Classically, regression models represent the dependence of a response variable in function to a set of predictor variables known as covariates, regressors, or independent variables. Regression analysis has become one of the most important and widely used statistical techniques, allowing the construction of mathematical models to explain possible relationships between the response variable and the different covariates. Overall, a model is a small-scale, abstract representation of reality which allows us to understand and describe a phenomena. The ultimate goal of the model is to predict or estimate the value of a variable taking into account the value of a set of known covariates.

Regression techniques are widely used on multiple disciplines to study relationships between different variables. In particular, regression analysis is a very useful tool for many biomedical studies, for example, to study risk factors, to explore prognostic patterns or to derive predictions for individual patients, among others. A literature search on Web of Science's (available) databases (WoS, *https://clarivate.com/products/web-of-science/*) for the single keyword: *"regression model"* shows more than 120000 publications in more than 100 research areas (including biology, environmental sciences, social sciences, economics, philology studies, or engineering practice, among many other areas). Notice that "publication" refers herein to articles, reviews, clinical trials, case reports, and books. This basic research highlights the importance of multidisciplinary teams in this type of studies. High quality results require a mix of statistical algorithms, computer sciences and domain knowledge. In general, the collaboration of statisticians and specialists from different areas is essential to understand each particular studies' background. Quoting Albert Einstein's words: "*The problem formulation is more essential than its own solution, which may simply be a mathematical or experimental skill*". When studying a real practical situation, we should be aware that having a model of the data is worthless if we are not able to draw conclusions from it. In this thesis, given the complexity of the presented models, special attention was paid to the clinical interpretation of the results obtained.

The earliest form of regression was exposed for the first time at the end of

the 19th century. Since then, it has become a very active research area in both theoretical, and empirical frameworks. The advances carried out to improve the flexibility in both predictor specifications, and response variable were numerous during last years.

Let us assume that observations $(y_i, i = 1, \ldots, n)$ are made, where y_i are observations on the response variable, and $(\nu_{1i}, \nu_{2i}, \ldots, \nu_{mi})$ represent all covariate information for individual i, for example binary, categorical or continuous effects. In this scenario, the simplest regression model assumes the relationship between the response and the covariates - plus random noise (ϵ) - to be linear

$$y_i = \beta_0 + \beta_1 \nu_{1i} + \cdots + \beta_m \nu_{mi} + \epsilon,$$

where $\beta_j, j = 0, \ldots, m$ are unknown regression coefficients which must be estimated. In linear regression models, statistical inference is based on the assumption that the response variable is normally distributed. The need to extend this type of regression models for a flexible approach - in order to consider other types of response variables such as binary, categorical, or other non-normal continuous outcomes - gave rise to the well known Generalized Linear Models (GLM, Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989).

GLM regression models allow for different distribution responses a part from the normal and for a degree of non linearity in the model structure. In this type of models, the (conditional) expectation $\mathbb{E}(Y_i \mid \nu_i) = \mu_i$ is linked to a linear predictor, η as follows

$$\eta_i = g(\mu_i) = \beta_0 + \beta_1 \nu_{1i} + \cdots + \beta_m \nu_{mi}, \tag{1.1}$$

where g is a known smooth link function[1] which ensure that the restrictions on the parameter spaces are maintained. GLM assumes that the response variable belongs to the exponential family distribution (e.g, Poisson, binomial, gamma or normal distribution among others). Note that linear regression models are a particular case of GLM in which the response variable follows a normal distribution and the link function is the identity.

Even though GLM are more flexible than classical linear models in terms of response variable distribution, the main limitation of this methodology is the assumption of linearity for the covariate effects. In most practical cases, it is sensible to assume that the goodness of fit will improve if the covariates follow an unknown non-linear function, see Figure 1.1.

Since the early 1990s, Generalized Additive Models (GAM, Hastie and Tibshirani, 1990) overcame the problems already discussed above. In GAM regression models the linear predictor in equation (1.1) is replaced by an additive predictor as follows

$$\eta_i = g(\mu_i) = \beta_0 + f_1(\nu_{1i}) + f_2(\nu_{2i}) + \cdots + f_m(\nu_{mi}), \tag{1.2}$$

[1] In the following, the inverse of this function will be denote by h.

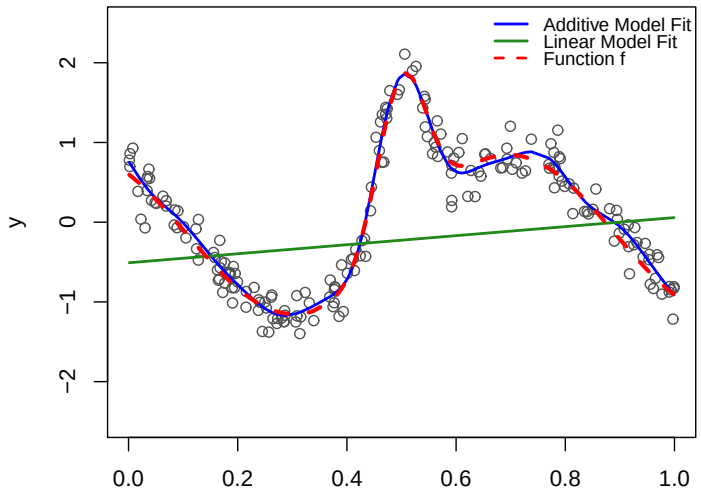


Figure 1.1: Differences between smooth (in blue) and linear models fitted (in green) for a simulated data where $\nu_1 \sim U\,[0,1]$ and $y = f(\nu_1) + \epsilon$, with $f(\nu_1) = \sin\left(2(4\nu_1 - 2)\right) + 2\exp\left(-16^2(\nu_1 - 0.5)^2\right)$ and $\epsilon \in N(0, 0.2)$.

where β_0 is a global intercept and $f_j, j = 1 \ldots m$ are unknown "smooth functions" which must be estimated. GAM models allowed for several covariate effects to be taken into account, including for example smooth estimation of the effect of continuous variables, categorical covariates, random effects, interactions, and possible spatial or temporal trends. Part of this flexibility comes from the introduction of the smooth functions. There are multiple smoothers that can be considered in practice (Wood, 2006), such as penalized splines (P-splines, Eilers and Marx, 1996; Lang and Brezger, 2004), thin plate regression splines (Wood, 2011) or Gaussian Markov random fields (Fahrmeir and Kneib, 2011), among others. This topic will be examined in more detail in Chapter 2.

GAM regression models have been widely explored in the scientific literature. In order to evaluate the impact that GAM regression models have produced in statistical and biomedical literature, we conducted an analysis of total number of citations. A global search on the databases of the WoS was carried out in order to have a rough appreciation of the impact of GAM models in the statistical and biomedical literature. For this aim, we have analysed the number of publications associated with *"GAM"* or *"Generalized Additive Model*"* or *"Generalized Additive regression Model*"*. Figure 1.2 shows the WoS's search results; as we can see, the number of citations of GAM regression models is growing over the years. We have found more than 7000 publications from 1990 to 2018; most of

the above papers are in the field of environmental sciences and biology (around 42%), mathematics (around 26%), and biomedicine (around 14%).

Inference in GAM regression models can be carried out from frequentist (Hastie and Tibshirani, 1990; Wood, 2006) and Bayesian inference through Structured Additive Regression models (STAR, Fahrmeir et al., 2013). STAR regression models were proposed in 2004 as a generalization of well-known model classes such as the GAM or geoadditive regression models (Kammann and Wand, 2003). However, nowadays both approaches allow for equivalent flexibility. Figure 1.2 shows the number of citations of STAR regression models in WoS with keywords: *"Structured additive regression"* or *"STAR regression models"*. Most of them are in the fields of statistics and probability, computational science applications, and economics.

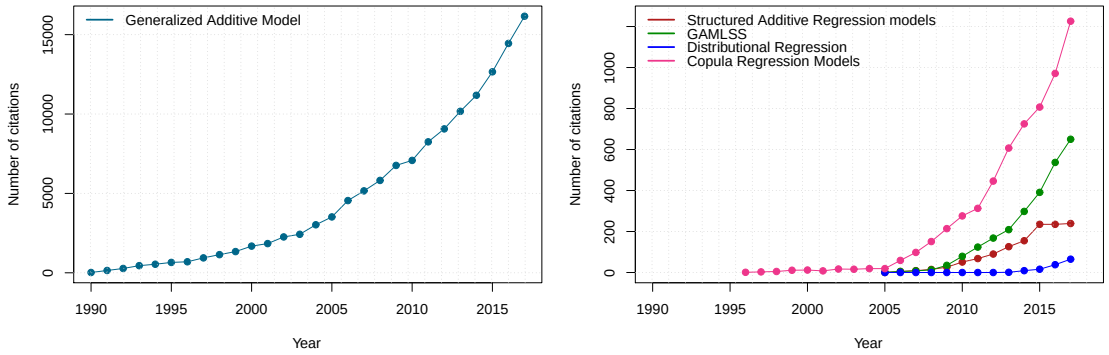


Figure 1.2: Number of citations per year between 1990 and 2018. Data retrieved from WoS.

However, while GAM and STAR are much more flexible than GLM, they still have some limitations. Although knowing the mean is very important in regression analysis, the interest in models beyond the mean has increased during recent years.

1.1 Regression beyond the mean: Distributional regression

As mentioned above, in classical regression models it is common to study the mean of the response variable as a function of the values of the explanatory variables. However, focusing solely on means may lead to an over-simplified picture of the situation. In fact, it is important for many applications to characterise the effects of covariates on all the parameters of the response's distribution, such as variance, or to know the complete distribution of the response (Espasandín-Domínguez et al., 2018b).

GAM regression models for Location, Scale and Shape (GAMLSS, Rigby and Stasinopoulos, 2005) overcome these difficulties. In the GAMLSS framework, the response, Y_i, is assumed to follow a parametric complex distribution, which does not have to be a member of the simple exponential family as in mean regression. Furthermore, in this type of regression model, every parameter of the response distribution - rather than just the mean - is related to an additive predictor.

Similar to GAM models, GAMLSS assumes a fully parametric specification for the distribution of the univariate response vector, $Y_i \mid\sim D(\mu_i, \sigma_i, \upsilon_i, \tau_i)$, where the parameters of the response distribution can be expressed as a function to the explanatory variables. Rigby and Stasinopoulos (2005), define GAMLSS regression models as follows:

$$\begin{cases} \eta^{\mu} = g_1(\mu) = \beta_0^{\mu} + f_1^{\mu}(\nu_{1i}) + f_2^{\mu}(\nu_{2i}) + \cdots + f_m^{\mu}(\nu_{mi}) \\ \eta^{\sigma} = g_2(\sigma) = \beta_0^{\sigma} + f_1^{\sigma}(\nu_{1i}) + f_2^{\sigma}(\nu_{2i}) + \cdots + f_m^{\sigma}(\nu_{mi}) \\ \eta^{\upsilon} = g_3(\upsilon) = \beta_0^{\upsilon} + f_1^{\upsilon}(\nu_{1i}) + f_2^{\upsilon}(\nu_{2i}) + \cdots + f_m^{\upsilon}(\nu_{mi}) \\ \eta^{\tau} = g_4(\tau) = \beta_0^{\tau} + f_1^{\tau}(\nu_{1i}) + f_2^{\tau}(\nu_{2i}) + \cdots + f_m^{\tau}(\nu_{mi}), \end{cases} \tag{1.3}$$

where the first equation (1.3) refers to the location parameter (μ) of y_i. The second refers to the scale parameter, σ, and the last two refers to the shape parameters (υ, τ).

A global search on the WoS' databases was carried out in order to evaluate the impact of these methodologies in the literature. For this aim, we searched for the following keywords: *"GAMLSS"* and/or *"Generalized Additive Model* for Location, Scale and Shape"*. Figure 1.2 shows the results of our literature search. We found a total of 182 publications with the above keywords. Most of the above works are in the field of mathematics (54.40%), paediatrics (around 26.67%), physiology (around 25.28%), meteorology sciences (21.43%) or environmental sciences (around 20.9%). Note that the same article can be categorized in various research areas simultaneously.

There are also alternatives to GAMLSS regression models working from a Bayesian framework, these are known as distributional regression (DR) models (Klein et al., 2014b, 2015). This latter type of models suppose an extension of the STAR models mentioned above. The notation of DR is more general than that of GAMLSS and takes into account the fact that, on many occasions, the estimated parameters of the response are not directly related to their localization, scale or shape, but to more general parameters.

Similar to the GAMLSS formulation, in DR, each parameter, $\vartheta_{i1}, \ldots, \vartheta_{iK}$, of the response distribution is related to an additive predictor $\eta_i^{\vartheta_k}$ defined in terms of the covariates. As in other types of classic regression models, a suitable response function is used to map the predictor to the parameter of interest, i.e.

$$\vartheta_{i1} = g_1(\eta_{i1}), \vartheta_{i2} = g_2(\eta_{i2}), \ldots, \vartheta_{iK} = g_K(\eta_{iK}). \tag{1.4}$$

The predictor of equation (1.4) can vary over different covariates by using additive predictors. Moreover, for an observation $i = 1, \ldots, n$, a suitable structured

additive predictor for parameter ϑ_k can be written as:

$$\eta_i^{\vartheta_k} = \beta_0 + f_1^{\vartheta_k}(\nu_{1i}) + \cdots + f_{J_k}^{\vartheta_k}(\nu_{J_k i}), \tag{1.5}$$

where the functions $f_j^{\vartheta_k},\ j = 1, \ldots, J_k$ represent the different covariate effects for each parameter of the response distribution.

In the present thesis, we will examine different distributions that depend on the mean and variance of responses, but in other studies involving other response variables it might be necessary to contemplate more complex distributions. It is the versatility of distributional regression models that allow them to incorporate complex distributions, such as the Dagum distribution, which depends on one scale and two shape parameters; the mean of the response is proportional to one of these parameters. Although the other parameters are not that easily interpretable at first sight, the Dagum distribution has the great advantage that both the conditional mode and conditional quantiles can be expressed in closed form. However, the interpretation of the results provided by distributional regression models is not always easy. This is the reason why Chapter 3 includes a guide to make correct interpretations of DR.

In this thesis, we will focus on the study of DR modelling and will present some extensions about this modern regression technique to consider multivariate responses and functional data covariates. Nowadays, technological progress has led to the development of new measurement procedures in the form of functional data, such as measurements over fine time or space grids and images with many pixels, (Febrero-Bande and Oviedo de la Fuente, 2012). There are multiple examples of functional data in medicine, e.g continuous glucose monitoring, data on electrical activity along the scalp (i.e. electroencephalography), data on electrical activity of the heart, among other scientific and industrial settings. Figure 1.3 shows a prototype for the type of data that we will consider. It shows the measurements of glucose levels for 8 people per 5 minutes. In this case, glucose levels are equally spaced, but in many other applications spacing measurements could be unequal.

The textbook by Ramsay and Silverman (1997) served as the starting point in the development of functional data methodologies, which has accelerated in the past decade to become one of the most promising areas in statistics. The number of applications yielding this type of data also support it.

In the framework of regression analysis, a regression model is known as "functional" if (at least) one of the involved variables (the dependent variable or some of the independent variables) are functional. In the statistical literature, there has also been a great deal of work in functional predictor regression or functional regression, especially in the past 10 years. Figure 1.4 shows the WoS search results with keywords "*signal regression*" and "*functional regression*" where we can observe the recent attention on regression models including functional data. A total of 801 publications have been found from 1995 to 2018, with a trend

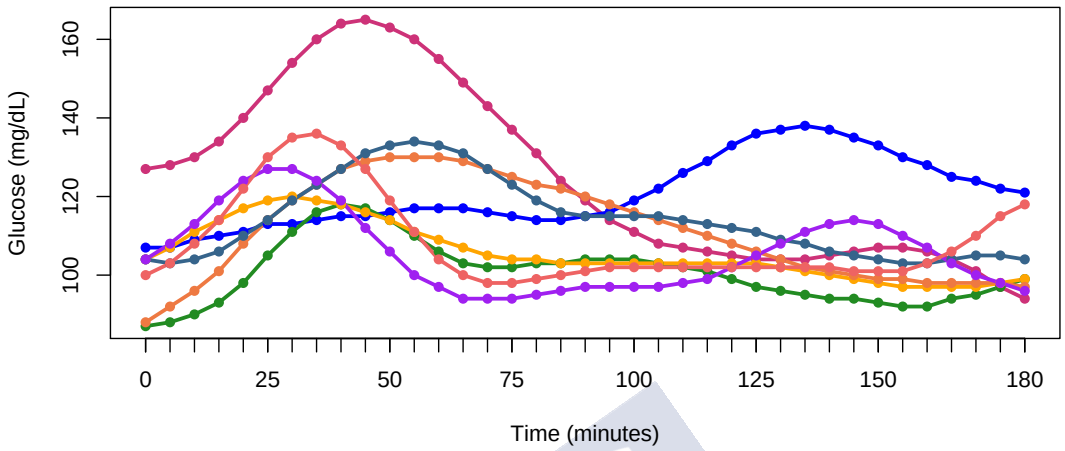


Figure 1.3: An example of functional data. *Spaghetti* plot for glucose measurements recorded for 8 random subjects.

which appears to be exponential. Most of these articles are in the field of statistics and probability (33.9%); approximately 28% of them make some application in medical research. The areas with more applications in biomedicine are neurology, radiology, psychology, and genetics.

In this thesis, we propose to incorporate this functional information within the framework of DR models. The methodologies herein developed will be applied to real biomedical data in a study of glycated proteins. The predictor will include the results of continuous monitoring, for which a *spaghetti* plot is shown in Figure 1.3 for 8 random subjects.

1.2 Copula distributional regression models

In biomedical applications, it is also often necessary to model jointly two or more responses as well as to determine the relationship between them. However, in most published regression studies for multivariate responses, a specific distribution is assumed for the response variable for no apparent reason and there are few contributions using non-parametric predictors. Moreover, flexible covariate effects are not typically considered (Espasandín-Domínguez et al., 2018b).

In recent years, different regression methodologies for bivariate responses based on copula functions have been developed in statistical literature. A major advantage of the copula approach is that marginal distributions may also come from different non-standard families (see, for example Marra and Radice, 2017a). However, most of the existing multivariate distributions are simple ex-

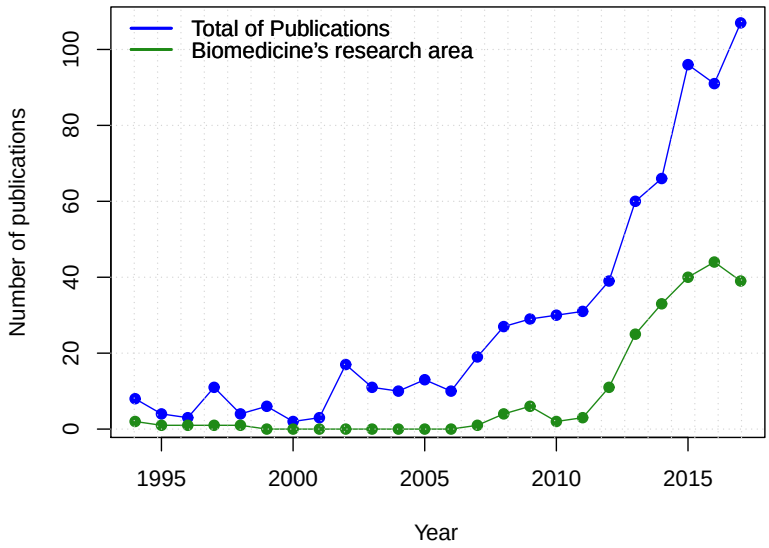


Figure 1.4: Number of publications per year including the keywords *"signal regression"* and *"functional regression"* between the years 1995 and 2018 retrieved from WoS.

tensions of the univariate distributions and often have the restrictive properties that all of the marginal distributions are of the same type (e.g., by construction, all marginal distributions of a normal multivariate are normal).

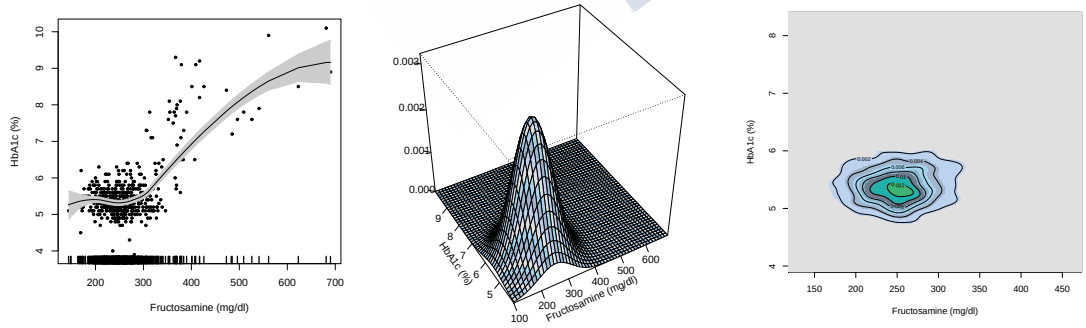


Figure 1.5: Relationship between fructosamine and glycated haemoglobin (HbA1c) (left). The middle panel is the perspective plot of the Kernel density estimate using a Gaussian kernel (see Bowman and Azzalini, 1997) from fructosamine and HbA1c. The right panel shows the same surface by contour plotting.

Dependence modelling using copula functions has become very popular in recent years as a multivariate modelling tool in many fields where multivariate

dependence is of interest and standard multivariate normality is in question. This methodology is particularly useful in the field of medicine. For instance, the biostatistical aim of this thesis is to investigate the discordances found among the results for two different glycated proteins (as glycated haemoglobin and fructosamine) useful in the diabetes diagnosis. (See Figure 1.5). But there are many other examples where multivariate regression models can be very useful. In fact, in many biomedical studies, it is necessary to model several (say d) response distributions: $(Y_1, Y_2, \ldots, Y_d)$, for example to study i) the same pathology in the eyes (to detect diabetic retinopathy, ii) the relationship between different diseases in the same patient (as diabetes or hypertension), iii) or the relationship of the systolic and diastolic blood pressure. Furthermore, in this type of models it is also interesting to asses the structure of dependence between different response variables, together with the effect of the covariates on this structure.

In this thesis, we will focus on copula distributional regression models in both frequentist (Marra and Radice, 2017a) and Bayesian frameworks (Klein and Kneib, 2016b). This novel approach extends the use of GAMLSS (Rigby and Stasinopoulos, 2005) and distributional regression (Klein et al., 2014a,b) to situations in which each parameter of a multivariate response is modelled simultaneously on some conditional covariates using different copula functions. Furthermore, this type of regression model enables the modelling of all distributional parameters using additive predictors that allow for several types of covariate effects, such as non-linear effects of continuous covariates, random effects or interactions (Espasandín-Domínguez et al., 2018b).

Figure 1.2 shows the citations that copula regression models have been received. A search in WoS using the following keywords *"copula additive model*"*, or *"copula regression"*, or *"copula model*"*, or *"CGAMLSS"*, or *"joint copula models"*, or *"copula additive regression"*, showed 548 manuscripts. An historical retrospective illustrates that copula regression models have been linked to mathematical (70%), business economics studies (37.23%), computer sciences (24.64%), or mathematical methods in social sciences (10.04%). However, less than 15 publications have been found in the framework of life sciences or biomedicine. Furthermore, most of these biomedical manuscripts are survival studies or publications in the field of mathematics where the developed methods were illustrated using environmental data (without any interpretation). In this thesis, we want to highlight that these modern regression techniques may be useful for clinicians since they allow for simultaneous explanation of the mechanisms affecting on multivariate responses; therefore, using these models could shed light on certain important biological processes. This methodology was proved to be useful in the setting of diabetes research (see Chapter 4).

1.3 General objectives of the thesis

The general aim of this thesis is to offer contributions in the distributional regression framework for both univariate and bivariate responses. Specifically, the main scientific contributions of this thesis, include the following:

i) To compare for the first time the different Bayesian and frequentist copula DR approaches, ii) to introduce functional data covariates in the framework of DR methodologies for one and two response models, iii) to provide a user's guide to interpret and visualize the results of DR models in real biomedical studies.

1.4 Structure of the thesis

Biostatistics and medicine are disciplines which share common goals in terms of improving the quality of life of the people through actions in the promotion of health and in the prevention of disease, understanding the disease as a result of the interaction between the individual and the environment.

Understanding and building knowledge from health data requires developing advanced analytical techniques that can transform data into meaningful information. The analytical solutions when applied to healthcare data have an immense potential to transform healthcare delivery. We all know that the biggest stimulus of new tools and theories of Data Science is the analysis of data to solve problems posed in terms of the subject matter under investigation. Creative researchers, faced with problems posed by data, will respond with a wealth of new ideas that often apply much more widely than the particular data sets that gave rise to the ideas.

The challenge addressed in this thesis presents many facets from a biostatistical standpoint, in that it calls for modern statistical techniques such as DR models, with extensions to methodologies such as joint modelling and functional data analysis, and other techniques that would be applicable, not only to the case studies introduced in this thesis, but also to other chronic and prevalent diseases. More specifically, in this thesis, a series of statistical techniques are developed trying to answer challenges which raised from a clinical study, that is AEGIS (A Estrada Glycation and Inflammation Study). In order to extend distributional regression techniques to the spatial environment, a second application related to clinical decision making is also presented, the blood potassium levels.

The present work is structured in eight chapters. This chapter introduces the motivation, objectives, and structure of the present thesis. In a nutshell, the objective of this work is to review and to provide new contributions in the framework of distributional regression for univariate and bivariate responses. Furthermore, this thesis is presented from the framework of Biomedical Data Science. In this sense, different techniques for the correct interpretation of the obtained results

will be exposed in this thesis. The remaining chapters of this work are structured as follows.

Chapter 2, introduces the theoretical background on distributional regression models. Chapter 3, proposes the use of reference bands in combination with quantile-quantile plots for determining the goodness of fit of a structured additive distributional regression. We adapted the methodology introduced by Augustin et al. (2012) to the context of distributional regression - for the first time. The behaviour of these graphics was tested via a simulation study. Furthermore, Chapter 3 illustrates how to visualize the results of distributional regression models in an analysis comprising spatial information, and the interpretation of results obtained using a novel distributional regression model. In this sense, we also provide the designed code to reproduce the analysis. Some of the contents of this chapter were published in 2018 by *Spatial Statistics* (Espasandín-Domínguez et al., 2018a).

Chapter 4, reviews novel approaches on bivariate copula distributional regression models and compare them via a simulation study and in a biomedical study. This chapter shows the usefulness of this type of models in practice and provides a basis for its clinical interpretation and reproduction by means of an understandable biomedical study. To the best of our knowledge, this is also the first time that copula regression have been used in the context of diabetes research. We also provide the programming code to reproduce the analysis.

Chapter 5, presents an extension of Brockhaus et al. (2018) methodologies, on how to include functional data covariates in a distributional regression model. This is the first time, that functional data covariates have been considered on this type of models. This extension has been evaluated through a simulation study. Furthermore, new programming code is provided to the user for the application of this extension on practice in an open software known as BayesX (Belitz et al., 2015, 2016). Chapter 6, presents current and future research on how to include functional covariates into CGAMLSS methodologies. In this chapter, the ideas of McLean et al. (2014) have been adapted to the frequentist CGAMLSS regression. The proposed technique has been used in a real biomedical study. Furthermore, new functions have been implemented in R-software (R Core Team, 2017), to adjust and interpret the proposed extension. Finally, Chapter 7 discusses the main results of the dissertation.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 2

Theoretical background on distributional regression

The interest of models beyond the mean has increased during the last years. Accordingly, in both frequentist (GAMLSS, Rigby and Stasinopoulos, 2005) and Bayesian frameworks (structured additive distributional regression models, Klein et al., 2015) new models beyond the mean have been developed as an alternative to classical regression. These types of models permit each parameter of the response distribution to be modelled using additive predictors that allow for several types of covariate effects.

This thesis will be developed in the framework of these distributional regression (DR) models. DR are semi-parametric regression models, i.e., DR have a "parametric" component (because a parametric distribution is assumed for the response variable) and a "nonparametric" part, because the parameters of the response distribution - as functions of explanatory variables - may involve nonparametric smoothing functions.

In this chapter, we will present some theoretical aspects on DR models. We will first introduce some preliminary concepts on flexible nonparametric regression techniques. Therefore, Section 2.1 provides an introduction to the flexible modelling of effects of continuous covariates on a dependent variable. Section 2.2 introduces the DR framework, and finally Section 2.3 summarizes possible inference procedures in DR available in the literature.

2.1 Basic concepts on nonparametric regression

Even though Generalized Linear Models may be sufficient for simple and linear relationships, they quickly become intractable in more complex situations, such as nonlinear relationships - as mentioned in Chapter 1. The main objective of nonparametric regression is the flexible modelling of effects of continuous covariates on a dependent variable (Fahrmeir et al., 2013).

In the following, we will examine several nonparametric approaches in both univariate (Section 2.1.1) and bivariate (Section 2.1.3) frameworks which allow flexible modelling of the effect.

2.1.1 Univariate smoothing

Let us assume that observations on the response y_i and the corresponding values of a continuous covariate ν_i are available for $i = 1, \ldots, n$ observational units.

The standard univariate nonparametric regression model assumes that the response variable can be explained through a deterministic function of the covariate plus an additive error term (ϵ), as follows (Fahrmeir et al., 2013)

$$y_i = f(\nu_i) + \epsilon_i, \;\; i = 1, \ldots, n. \tag{2.1}$$

As in other types of classic linear regression model, the errors are assumed to be independent and identically distributed with $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2, i = 1, \ldots, n$. It therefore follows that $\mathbb{E}(y_i) = f(\nu_i)$, and $\mathrm{Var}(y_i) = \sigma^2, \; i = 1, \ldots, n$.

Note that f is a generic function of ν_i, the different assumptions made for this function give rise to different modelling possibilities. Several methods have been proposed in the literature to estimate function f. In the following, we describe some of them.

Polynomial splines

First, we introduced polynomial splines. This approach is closely related to the idea of polynomial regression modelling. In a classical polynomial regression model, the effect of the covariate, on the response, is assumed to be a polynomial of degree l, as follows (Fahrmeir et al., 2013)

$$f(\nu_i) = \gamma_0 + \gamma_1 \nu_i + \gamma_2 \nu_i^2 \ldots + \gamma_{l-1} \nu_i^{l-1} + \gamma_l \nu_i^l.$$

In this particular setting, the vector of regression coefficients of the polynomials (denoted by $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_{l-1}, \gamma_l)$ can be estimated using ordinary least squares. However, it is well known that, in general, a polynomial model is not flexible enough to capture nonlinear patters of the covariates. See Figure 2.1 for an illustrative example where a third order polynomial does not provide a good fit. A possible solution could be to increase the order of the polynomial, but this solution does not necessarily work, which might be due to a possibly different behaviour in different ranges of the covariate. As an alternative, we can split the domain of the covariates into several intervals and fit different polynomials in each interval (Fahrmeir et al., 2013). Figure 2.2 shows an example of this approach known as piecewise polynomial regression. However, as we can see the estimation obtained is not an overall smooth function and they show different values at the interval boundaries (see left panel of Figure 2.2). In contrast,

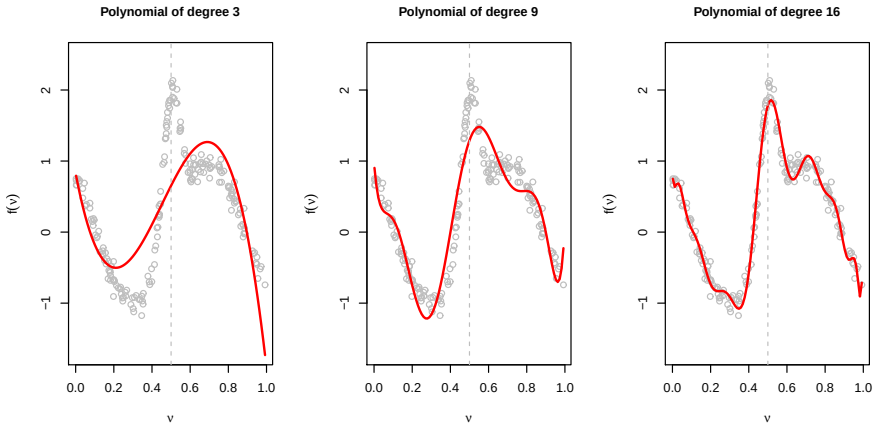


Figure 2.1: Different polynomial regression models for a simulated dataset. The data has been simulated according to the model $y = f(\nu) + \epsilon$ with $f(\nu) = \sin(8\nu - 3) + 2\exp -256(\nu - 0.5)^2)$ and $\epsilon \sim N(0, 0.09)$. In this example, we have considered polynomials of different degrees but they are not able to explain the data (either because they are too smooth or because the estimation is wiggly).

it would be desirable to obtain a function similar to the one shown in the right panel of Figure 2.2. For this aim, we can impose several smoothness restrictions at the boundaries of f. This idea leads to the class of polynomial splines (Fahrmeir et al., 2013).

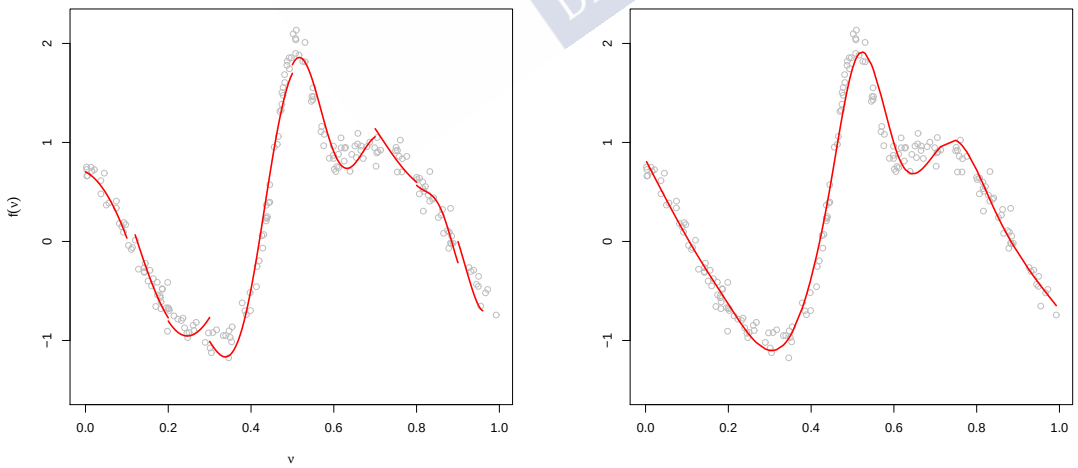


Figure 2.2: Examples of piecewise polynomial regression (left) and polynomial splines (right) for the simulated dataset introduced in Figure 2.1. Adapted from Fahrmeir et al. (2013) with permission.

A polynomial spline of degree $l \geq 0$ with knots, $a = k_0 < k_1 < \ldots < k_{m-1} <$

$k_m = b$, is a function $f : [a, b] \to \mathbb{R}$ that verifies the following conditions (Fahrmeir et al., 2013)

(i) $f(\nu)$ is $(l-1)$-times continuously differentiable. Note that $l = 0$ not need any smoothness requirement. The special case of $l = 1$, corresponds to $f(\nu)$ being continuous (but not differentiable). This condition ensures the desired smoothness restriction at the knots.

(ii) $f(\nu)$ is a polynomial of degree l for $\nu \in [k_j, k_{j+1})$, $j = 1, \ldots, m-1$.

Note that the degree of the spline refers to the global smoothness of the polynomial spline; the more number of knots considered, the higher the number of piecewise polynomials used. In the following, we will discuss the importance of the spline and degree as well as the selection of the knots. Before using polynomial splines, we need to define a corresponding regression basis. This basis can be achieved with different approaches. In the following, we will present two different ways of representation of the set of polynomial splines: the truncated power series, and B-splines.

Truncated power series

Let consider the following regression model:

$$y_i = \gamma_1 + \gamma_2 \nu_i + \cdots + \gamma_{l+1} \nu_i^l + \gamma_{l+2} (\nu_i - k_2)_+^l + \cdots + \gamma_{l+m-1} (\nu_i - k_{m-1})_+^l + \epsilon_i,$$

where

$$(\nu - k_j)_+^l = \begin{cases} (\nu - k_j)^l & \text{if} \quad \nu \geq k_j \\ 0 & \text{otherwise.} \end{cases}$$

The first part of the above model is a polynomial of degree l, while the rest of the coefficients change at every inner knot $k_2, \ldots, k_{m-1}$. This approach allows the use of local polynomials in every interval defined by the knots, moreover the global smoothness is ensured. See Figure 2.3 and Fahrmeir et al. (2013) for a more detailed illustration of the concept of a polynomial spline.

Theoretically, it can be shown that each polynomial spline of degree l with knots $k_1, \ldots, k_m$ can be expressed as a linear combination of the $d = m + l - 1$ basis functions:

$$\left\{ \begin{array}{l} B_1(\nu) = 1, B_2(\nu) = \nu, \ldots, B_{l+1}(\nu) = \nu^l, \\ B_{l+2}(\nu) = (\nu - k_2)_+^l, \ldots, B_d(\nu) = (\nu - k_{m-1})_+^l. \end{array} \right\} \tag{2.2}$$

Following this notation, equation (2.1) can be rewritten as

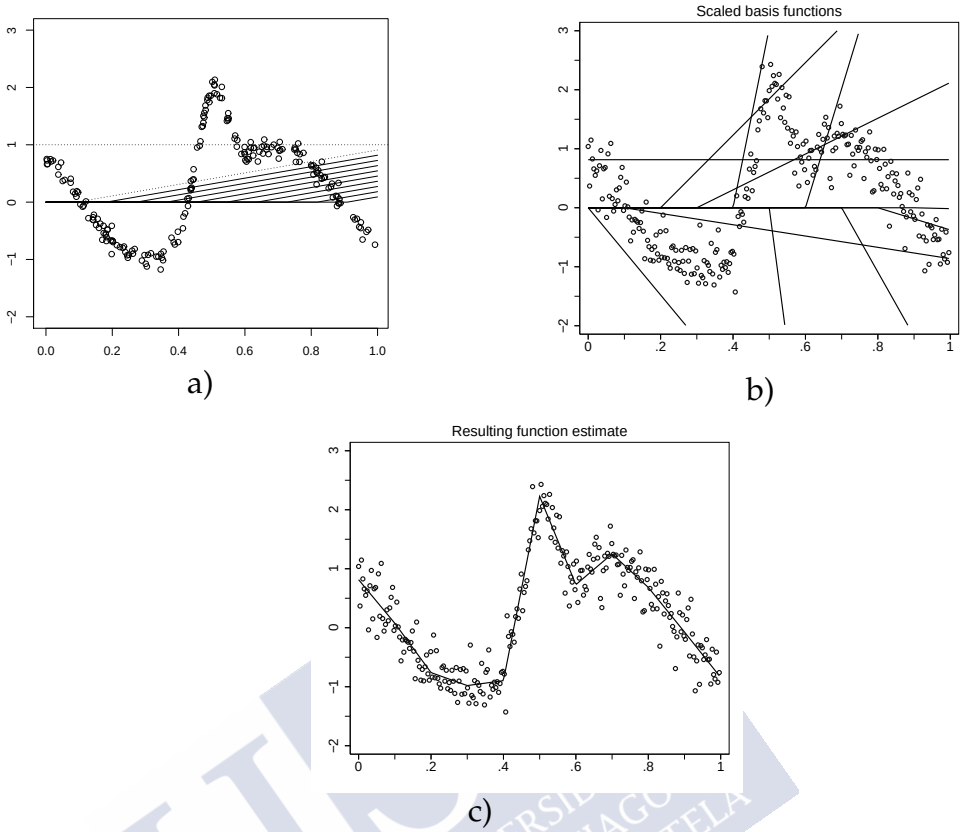


Figure 2.3: Illustration of a polynomial spline fit with linear truncated polynomials. In these panels are represented the basis functions (a), the scaled basis functions (b) and, the sum of the scaled basis functions (c). Adapted from Fahrmeir et al. (2013), with permission.

$$y_i = f(\nu_i) + \epsilon_i = \sum_{j=1}^{d} \gamma_j B_j(\nu_i) + \epsilon_i. \tag{2.3}$$

The functions $\{B_j,\ j = 1, \ldots, d\}$ are called basis functions because they allow to represent all polynomial splines. Different types of basis can be used for this pursuit. The basis defined in (2.2) are known as truncated power series basis (TP basis).

The mathematical modelling of $f(\nu)$ as a polynomial spline has the advantage that regression modelling can be understood as a linear model. However, the number of parameters considered is (possible) large.

Let be $\boldsymbol{y}$ the vector of observed response variable, $\boldsymbol{\epsilon}$ the vector of errors, and

$\boldsymbol{Z}$ the design matrix defined as

$$\boldsymbol{Z} = \begin{pmatrix} B_1(\nu_1) & \dots & B_d(\nu_1) \\ \vdots & & \vdots \\ B_1(\nu_n) & \dots & B_d(\nu_n) \end{pmatrix} = \begin{pmatrix} 1 & \nu_1 & \dots & \nu_1^l & (\nu_1 - k_2)_+^l & \dots & (\nu_1 - k_{m-1})_+^l \\ \vdots & & & & & & \vdots \\ 1 & \nu_n & \dots & \nu_n^l & (\nu_n - k_2)_+^l & \dots & (\nu_n - k_{m-1})_+^l \end{pmatrix}.$$

Equation (2.3) can then be rewritten as a linear model with regression coefficients $\boldsymbol{\gamma}$ as follows

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_d)'$ is the vector of coefficients. The defined linear model can be estimated using usual least squares

$$\widehat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}.$$

It should be noted that the regression coefficients can not be interpreted as in linear models. In this case the quality of the model can be checked in a scatterplot of the data, using the estimated curve (see Fahrmeir et al., 2013). The main choices that must be made here are the number and position of the knots, and the degree of the splines. Splines of degree 3, (cubic splines) are often used as default, because they lead to smooth and twice continuously differentiable function. In addition to the number of knots, we must decide the position of knots along the covariate axis. In practice, it is common to define i) equidistant knots; ii) quantile-based knots; or iii) to select visually the best location of the knots based on a scatter plot (see Fahrmeir et al., 2013). However, the major disadvantage of this approach is to determine the optimal number of knots to consider. As expected, when the number of knots is too large, the estimation is wiggly, and when the number of knots is too small, the fit looks over-smooth. To overcome these problems, there are basically two alternatives i) to introduce a penalty; or ii) the automatic data-driven selection based on model choice strategies (see Section 8.1.10 of Fahrmeir et al., 2013, for a detailed discussion). The first option will be considered here.

B-splines

Alternative bases to the TP basis in the context of polynomial splines have been proposed with better numerically properties as the B-splines basis. The main advantage of B-splines is their local definition. Furthermore, B-spline basis have only positive values on an interval based on $l + 2$ knots and they are bounded; (note that TP basis can lead to numerical instabilities for covariates with large values due to their construction from truncated polynomials).

A B-spline of order $l = 0$ (see Figure 2.4) is defined as

$$B_j^0(\nu) = I(k_j \leq \nu < k_{j+1}) = \begin{cases} 1 & k_j \leq \nu < k_{j+1},\ j = 1, \dots, d-1 \\ 0 & \text{otherwise,} \end{cases}$$

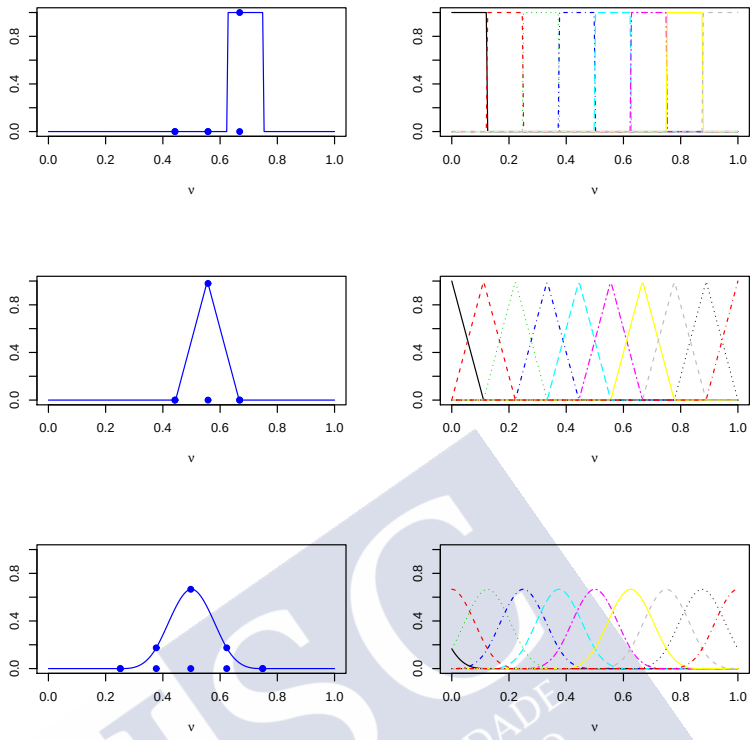


Figure 2.4: B-spline basis functions of degrees $l = 0$ (top), 1 (middle) and 3 (bottom) using equidistant knots.

where $I(.)$ denotes the indicator function. B-splines of higher order can be defined recursively. A B-spline of order $l \geq 1$ is defined as follows

$$B_j^l(\nu) = \frac{\nu - k_{j-l}}{k_j - k_{j-l}} B_{j-1}^{l-1}(\nu) + \frac{k_{j+1} - \nu}{k_{j+1} - k_{j+1-l}} B_j^{l-1}(\nu).$$

Figure 2.4 shows several examples of B-spline basis functions for degrees $l = 0, 1$ and, 3 considering equidistant knots.

In general, a B-spline of degree l has the following characteristics (Durbán, 2009)

(i) It consists of $l + 1$ pieces of polynomials of degree l joined by l inner knots.

(ii) As each B-spline is composed by set of polynomials, it is easy to calculate its derivative. Specifically, the derivative of each basis function can be expressed as follows

$$\frac{\partial}{\partial \nu} B_j^l(\nu) = l \left(\frac{1}{k_j - k_{j-1}} B_{j-1}^{l-1}(\nu) - \frac{1}{k_{j+1} - k_{j+1-l}} B_j^{l-1}(\nu) \right). \quad (2.4)$$

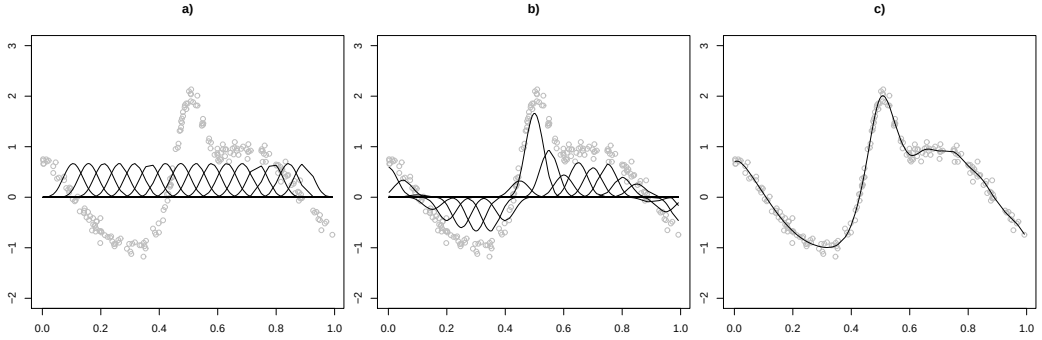


Figure 2.5: Illustration of a nonparametric estimation using cubic B-splines. First step is to compute the B-splines basis (a). Second step is to scale the B-spline basis (b) and finally to represent the sum of scaled B-spline basis functions (c). Adapted from Fahrmeir et al. (2013) with permission.

From this equation (2.4), we can obtain the derivative for the entire polynomial spline as follows

$$\frac{\partial}{\partial \nu} \sum_j \gamma_j B_j^l(\nu) = l \sum_j \frac{\gamma_j - \gamma_{j-1}}{k_j - k_{j-l}} B_{j-1}^{l-1}(\nu). \tag{2.5}$$

(iii) Moreover, for every single basis function the $(l-1)$-derivative is continuous on the connecting points.

(iv) It is positive on its expanded domain by $l+2$ adjacent knots and 0 on the rest of the points.

(v) For every point $\nu \in [a, b]$, we have $\sum_{j=1}^{d} B_j(\nu) = 1$.

(vi) Every basis function, within the domain $[a, b]$ overlaps with exactly $2l$ adjacent basis functions (except at the extremes).

(vii) For every point of ν, we have that $l+1$ B-splines are non zero.

Similar to the TP basis, we can define the design matrix as follows

$$\boldsymbol{Z} = \begin{pmatrix} B_1^l(\nu_1) & \dots & B_d^l(\nu_1) \\ \vdots & & \vdots \\ B_1^l(\nu_n) & \dots & B_d^l(\nu_n) \end{pmatrix}.$$

Note that the rows of matrix $\boldsymbol{Z}$ sum to one (see property iv). The defined matrix, $\boldsymbol{Z}$, does not contain an explicit intercept term. Specifying an additional intercept would lead to an unidentifiable model. Furthermore, matrix $\boldsymbol{Z}$ mainly consists of zeros due to the definition of B-splines. In this framework, the normal equation

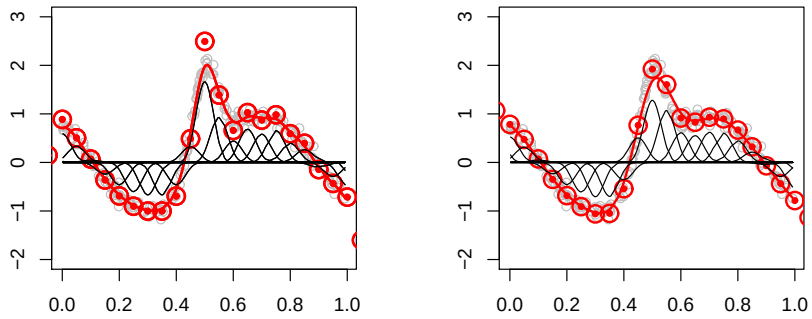


Figure 2.6: Comparison of a nonparametric fit using cubic B-splines with 20 knows (left) and P-splines (right). Adapted from *"An introduction to smoothing with penalties: P-splines"* by M. Durbán, Boletín de Estadística e Investigación Operativa 2009, 25, p. 201. Copyright 2009 by SEIO.

$\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\gamma} = \boldsymbol{Z}'\boldsymbol{y}$ can be solved in a numerically efficient way. Figure 2.5 illustrates the estimation of a B-spline fit for the simulated data considered at the beginning of the chapter.

Though B-splines solve possible numerical instabilities and collinearity problems of TP basis, B-splines also depends on the number of knots. A possible solution to overcome this problem is to introduce penalizations, as mentioned before. In the following section, we will summarize the main idea of penalized splines.

Penalized splines (P-splines)

The idea of P-splines can be summarized in the two following steps (Fahrmeir et al., 2013)

(i) To approximate the function $f(\nu)$ with a polynomial spline using a "sufficient" number of knots. The placement of the knots has only a very minor impact on the fit if the number of knots chosen is not too small. In general, 20 equidistant knots yield sufficient flexibility for basically all situations of applied interest.

(ii) To introduce an additional penalty term that prevents over-fitting and minimize a penalized least squares instead of the common least squares criterion. (Eilers and Marx, 1996; Lang and Brezger, 2004; Brezger and Lang, 2006). See also Figure 2.6 and Durbán (2009).

The main advantage of the penalization is that the smoothness not depends on the number and the position of knots but rather by one smoothing parameter.

In the remainder of the chapter, equidistant knots will be considered to simplify the notation.

In the same way as B-splines, P-splines can be based on a TP or B-splines, among other types of basis. In this thesis, we choose the representation through B-splines.

P-splines based on a B-spline basis

If $f(\nu)$ is represented using a B-spline basis, it can be shown that the use of (integrated squared) derivatives is an appropriate penalty because they represent the variability of the function (Eilers and Marx, 1996; Wood, 2006; Fahrmeir et al., 2013). A popular penalty based on the second derivative is given by

$$\lambda \int \left(f''(\nu)\right)^2 d\nu,$$

because it takes into account the curvature of the function. As can be seen in equation (2.5), the first derivative of a B-spline can be expressed in terms of the the differences between the regression coefficients. To avoid very large values of this derivative, penalties based on these differences can be introduced to obtain smooth functions. In addition, instead of obtaining a smooth function in terms of the first derivative, we can obtain a smooth function in terms of the r-th-order derivatives using differences of order r. Moreover, in a P-spline approach, the *penalized least sum squares* (PLS), is minimized instead of the classical *residual sum squares*. This PLS criteria is defined as follows

$$PLS(\lambda) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j(\nu_i) \right)^2 + \lambda \sum_{j=r+1}^{d} (\Delta^r \gamma_j)^2, \tag{2.6}$$

where $LS = \sum_{i=1}^{n} \left(y_i - f(\nu_i)\right)^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j(\nu_i) \right)^2$ denotes the classical *residual sum squares*. The smoothing parameter of equation (2.6), λ, controls the compromise between smoothness and fidelity to the data. For a large value of λ, the estimation of f is close to a polynomial of degree $r-1$, meanwhile when $\lambda = 0$, no penalty is considered (see Figure 2.7). The operator Δ^r defined in equation (2.6) represents the $r-$th-order difference applied to the B-spline coefficients. These operator is defined recursively as follows

$$\begin{aligned}
&\Delta^1 \gamma_j = \gamma_j - \gamma_{j-1} \\
&\Delta^2 \gamma_j = \Delta^1(\Delta^1 \gamma_j) = \Delta^1 \gamma_j - \Delta^1 \gamma_{j-1} = \gamma_j - 2\gamma_{j-1} + \gamma_{j-2} \\
&\quad \vdots \\
&\Delta^r \gamma_j = \Delta^{r-1}(\Delta^{r-1} \gamma_j) = \Delta^{r-1} \gamma_j - \Delta^{r-1} \gamma_{j-1}.
\end{aligned}$$

The penalty term defined in (2.6) can be also rewritten in matrix notation as follows

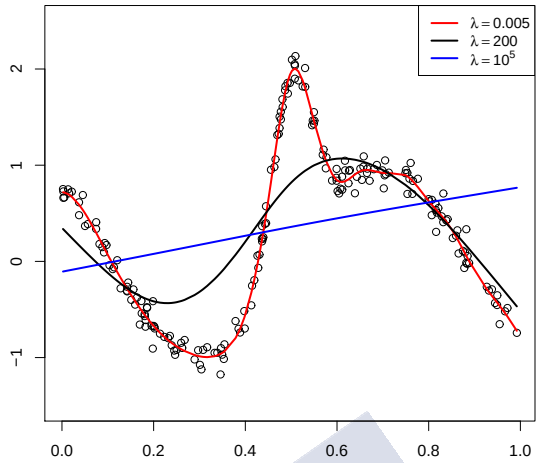


Figure 2.7: Impact of λ on cubic P-spline fit.

$$\lambda \sum_{j=r+1}^{d} (\Delta^r \gamma_j)^2 = \lambda \boldsymbol{\gamma}' \mathbf{D}_r' \boldsymbol{D}_r \boldsymbol{\gamma}, \tag{2.7}$$

where

$$\boldsymbol{D}_1 = \begin{pmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & \ddots & \ddots \\ & & & -1 \quad 1 \end{pmatrix} \in M_{(d-1)\times d},$$

and $\boldsymbol{D}_r = \boldsymbol{D}_1 \boldsymbol{D}_{r-1}, r \geq 2$ are difference matrices. Let denotes by $\boldsymbol{K}_r = \mathbf{D}_r' \boldsymbol{D}_r$, the difference penalty matrix, then equation (2.7) can be rewritten as

$$\lambda \sum_{j=r+1}^{d} (\Delta^r \gamma_j)^2 = \lambda \boldsymbol{\gamma}' \mathbf{D}_r' \boldsymbol{D}_r \boldsymbol{\gamma} = \lambda \boldsymbol{\gamma}' \boldsymbol{K}_r \boldsymbol{\gamma}.$$

Bayesian P-splines

Penalized splines can also be derived in a Bayesian framework (Lang and Brezger, 2004). In the following paragraphs, we will summarize Bayesian P-splines based on B-splines.

Let us start again from model (2.1)

$$y_i = f(\nu_i) + \epsilon_i = \sum_{j=1}^{d} \gamma_j B_j(\nu_i) + \epsilon_i, \tag{2.8}$$

with B-spline basis function B_j. In a Bayesian framework, instead of imposing a penalty, an appropriate prior assumption for $\boldsymbol{\gamma}$ will be supposed.

In the Bayesian framework, penalty differences are defined as random walks of order r (RW-r). A first order random walk (RW-1) for equidistant knots can be defined as (Fahrmeir et al., 2013)

$$\gamma_j = \gamma_{j-1} + u_j, \;\; u_j \sim \mathrm{N}(0, \tau^2), \;\; j = 2, \ldots, d,$$

i.e,

$$\gamma_j - \gamma_{j-1} = u_j, \;\; u_j \sim \mathrm{N}(0, \tau^2), \;\; j = 2, \ldots, d.$$

The above expression shows the relationship between the random walk and the first-order difference penalty. In this framework, we assume a non-informative prior distribution for γ_1 such that $p(\gamma_1) \propto const$. The error variance, τ^2, can be interpreted as an inverse smoothing parameter: the larger the variance, the larger the possible deviation from the conditional expectation. When only very little deviations from γ_j and γ_{j-1} are allowed, which means that the variance of the RW-1 is almost zero, it results in a constant trend of the sequence, $\gamma_1, \ldots, \gamma_d$. In contrast, when having a large variance, τ^2, neighbouring coefficients are able to deviate from each other, leading to a rough estimated function.

For random walks of higher order, analogous results can be derived. For example, the second order random walk (RW-2) for $\boldsymbol{\gamma}$, is defined by (Fahrmeir et al., 2013)

$$\gamma_j = 2\gamma_{j-1} - \gamma_{j-2} + u_j, \;\; u_j \sim \mathrm{N}(0, \tau^2), \;\; j = 3, \ldots, d,$$

or equivalently,

$$\gamma_j - 2\gamma_{j-1} - \gamma_{j-2} = u_j, \;\; u_j \sim \mathrm{N}(0, \tau^2), \;\; j = 3, \ldots, d,$$

assuming $p(\gamma_1) \propto const$ and $p(\gamma_2) \propto const$.

On the other hand, the joint distribution of the regression parameters, $\boldsymbol{\gamma}$, can be written in the general form of a multivariate but improper Gaussian distribution as follows

$$p(\boldsymbol{\gamma} \mid \tau^2) \propto \left(\frac{1}{2\pi\tau^2}\right)^{rk(\boldsymbol{K}_2)/2} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}'\boldsymbol{K}_2\boldsymbol{\gamma}\right), \tag{2.9}$$

with $\boldsymbol{K}_2 = \boldsymbol{D}_2'\boldsymbol{D}_2$, i.e.,

$$\boldsymbol{K}_2 = \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & & \\ & & 1 & -4 & -6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix}.$$

The precision matrix $\boldsymbol{K}_2$ is not full rank, having rank $d-2$. In general, the penalty matrix constructed from a random walk of order r has rank $d-r$, leading to an improper joint prior of $\boldsymbol{\gamma}$.

2.1.2 Other smoothing approaches

Section 2.1.1 has shown the study of the effect of one continuous covariate on the response in a nonparametric framework. In many biomedical applications, it is worthwhile to model not only the effect of one continuous covariate on the response but also of other covariates. The methodologies introduced can be easily extended to consider more than one covariate. See Fahrmeir et al. (2013) for more details. Sometimes the study of the interaction between two covariates is also needed, or the modelling, for example, of spatial effects (see Chapter 3), functional data (see Chapters 5 and 6), among others types of effects. Section 2.1.3 describes some of them.

2.1.3 Bivariate smoothing

An interaction between two covariates exists, if the effect of a covariate depends on the value of at least one other covariate (Fahrmeir et al., 2013). In the framework of a classical linear model, interactions between two categorical covariates or one continuous and one categorical variable can be modelled using dummy coding - assuming in the last case that the main effect of the continuous variable and the interaction effect is linear. When estimating interactions between two continuous covariates, the concept of *basis functions* - defined in the Section 2.1.1 - can be extended to the context of bivariate interaction surfaces through the use of tensor product basis.

Tensor product basis

Let $f(\nu_1, \nu_2)$ be a two-dimensional surface, in the following, we will summarize how to model the effect of $f(\nu_1, \nu_2)$, where ν_1 and ν_2 can be denote two continuous covariates, as well as, coordinates in a spatial model (Fahrmeir and Kneib, 2011). Let $\left\{B_j^{(1)}(\nu_1), j = 1, \ldots, d_1,\right\}$ and $\left\{B_s^{(2)}(\nu_2), s = 1, \ldots, d_2\right\}$ be the basis functions for ν_1 and ν_2, respectively. The tensor product basis consists of the products of all basis functions, i.e.

$$B_{js}(\nu_1, \nu_2) = B_j^{(1)}(\nu_1) \cdot B_s^{(2)}(\nu_2), j = 1, \ldots, d_1, s = 1, \ldots, d_2.$$

This definition leads to the following representation for $f(\nu_1, \nu_2)$:

$$f(\nu_1, \nu_2) = \sum_{j=1}^{d_1} \sum_{s=1}^{d_2} \gamma_{js} B_{js}(\nu_1, \nu_2).$$

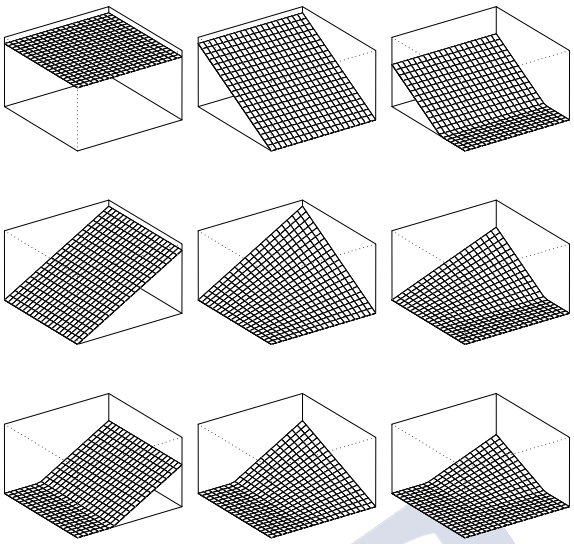


Figure 2.8: Illustration of a tensor product basis obtained from univariate linear TP basis. The first row and the first column were obtained by multiplying the constant basis functions in the direction of ν_1 with the ones of ν_2 direction. The remaining four basis correspond to the products of the rest of the basis. Adapted from Fahrmeir and Kneib (2011) with permission.

In the polynomial splines framework, we refer to the tensor product basis as tensor product splines or bivariate polynomial splines (Fahrmeir and Kneib, 2011). As an example, the tensor product spline based on the univariate TP basis functions: $\left\{B_1^{(1)}(\nu_1) = 1, B_2^{(1)}(\nu_1) = \nu_1, B_3^{(1)}(\nu_1) = (\nu_1 - k_1)_+\right\}$ and $\left\{B_1^{(2)}(\nu_2) = 1, B_2^{(2)}(\nu_2) = \nu_2, B_3^{(2)}(\nu_2) = (\nu_2 - k_2)_+\right\}$ is illustrated in Figure 2.8.

As in the univariate setting, we can use penalties to overcome the determination of the optimal number and position of knots. However the numerical difficulties - discussed in Section 2.1.1 - of TP basis increase here, and the use of B-spline basis is more appropriate. Figure 2.9 shows tensor product basis obtained from univariate B-splines of degrees $l = 0, 1, 2,$ and 3 respectively. (See for example: Dierckx, 1995; Fahrmeir and Kneib, 2011; Fahrmeir et al., 2013, for more details).

Note that the ideas developed in Section 2.1.1 can also be generalized for modelling higher dimensional surfaces. However, the number of parameters involved becomes to larger and this is the reason that in practice, it is common to assume an additive structure for a general function as follows

$$f(\nu_1, \ldots, \nu_m) = f_1(\nu_1) + \cdots + f_m(\nu_m).$$

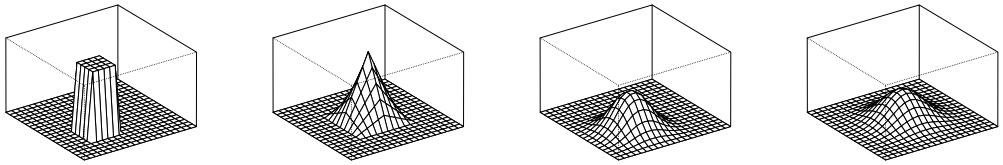


Figure 2.9: Tensor product basis obtained from univariate B-splines of degrees $l = 0, 1, 2,$ and 3 respectively. Note that the higher the l, the smoother estimation gets. Adapted from Fahrmeir and Kneib (2011) with permission.

2.1.4 Spatial smoothing

Spatial data structures

Spatial information can be collected on a continuous scale (as coordinates) or as discrete information. When considering continuous spatial information, the observations $\boldsymbol{s} = (s_1, \ldots, s_d), \boldsymbol{s} \in \mathbb{R}^d$ can be understood as a continuous variable (Fahrmeir and Kneib, 2011). An example of such continuous information could be the exact location of the residence of the patients in terms of coordinates (longitude and latitude). However, in many biomedical studies, due to practical or confidential reasons, it is not possible to know the exactly coordinate location and only discrete information is available. Such discrete information could be, e.g., a region, a district or a residence country. An example of this type of spatial information is the Potassium dataset - presented in Chapter 3. This dataset contains information of the place of residence aggregated by district on 145960 individuals of Health Area of Santiago de Compostela. Given the nature of these data, Section 3.3.2 of Chapter 3 discusses a common spatial smoothing approach based on Markov random fields.

It is also worth mentioning that although continuous and discrete spatial information are two different concepts; the discrimination between them is not easy in practice. Continuous information can also be treated within the framework of Markov random fields; for example by defining neighbours based on distance measures. On the other hand, discrete location variables can also be turned to into coordinate information by considering for example the centroids of the districts. This alternative, could be useful when the number of regions considered is large (Fahrmeir and Kneib, 2011).

Section 2.2 describes distributional regression models. Among other important properties, these models are able to incorporate spatial effects capturing unobserved spatial heterogeneity and spatial correlations. More details of this approach are shown in Chapter 3.

2.2 Distributional regression models

Distributional regression (DR) models are similar to quantile (Koenker and Bassett, 1978; Koenker and Ng, 2005) or expectile regression (Newey and Powell, 1987; Schnabel and Eilers, 2009; Sobotka and Kneib, 2012) in the sense that these types of models generalize classical mean regression models. However, there are several differences compared to quantile regression: i) DR also deals with discrete and mixed responses (including zero-inflated or over-dispersion) while quantile regression becomes less appropriate for these types of responses; ii) in DR framework, all parameters of the response distribution are estimated simultaneously, while in quantile regression, the different quantiles are estimated separately (yielding quantile crossing); iii) distributional regression models are semi-parametric regression models where a parametric distribution for the response variable is assumed.

Distributional regression models (as defined in Klein et al., 2014a) are innovative models that permit marginal distribution parameters to be modelled using additive predictors that allow for several types of covariate effects (such as the non-linear effects of continuous covariates, random effects, and the interactions or spatial effects). By modelling each parameter of the response at the same time - and not just the mean - they provide additional flexibility.

This type of model also provides different types of (possibly non-standard) response distributions for continuous, discrete, and mixed discrete continuous distributions. However, the adequate selection of response variable in the formulation of these models is not without its difficulty. See Chapter 3.

Distributional regression allows the effect of the covariate information on all the parameters of the response distribution to be examined.

2.3 Inference in distributional regression

Different inferential procedures on DR have been proposed in the statistical literature based on different techniques, which are summarised in the following diagram.

DR Inference Procedures in the literature

- i) Optimisation of the penalized log-likelihood (GAMLSS, Rigby and Stasinopoulos, 2005),
- ii) A Gradient Boosting approach (Mayr et al., 2012a),
- iii) Bayesian inference based on Markov Chain Monte Carlo (MCMC) (Klein et al., 2014a).

The penalized log-likelihood (denoted as l_p) of distributional regression mo-

dels is defined as follows (e.g. Rigby and Stasinopoulos, 2005; Klein et al., 2014a)

$$l_p = l - \frac{1}{2}\sum_{k=1}^{K}\sum_{j=1}^{J_k}\lambda_{j,k}\boldsymbol{\beta}'_{j,k}\boldsymbol{K}_{j,k}\boldsymbol{\beta}_{j,k}, \tag{2.10}$$

where $\boldsymbol{\beta}$ represents the vector of regression coefficients, $\boldsymbol{K}$ denotes the penalty matrix, λ is the smoothing parameter, K is the number of distributional parameters of the response distribution, and l denotes the log-likelihood defined as

$$\sum_{i=1}^{n}\log\left(p(y_i \mid (\vartheta_{i1},\ldots,\vartheta_{iK}))\right).$$

In the above equation, vector $(\vartheta_{i1},\ldots,\vartheta_{iK})$ represents the parameters of the response distribution.

2.3.1 Frequentist inference: Generalized Additive Models for Location Scale and Shape (GAMLSS)

Generalized Additive Models for Location, Scale and Shape (GAMLSS, Rigby and Stasinopoulos, 2005) are available in the R package called `gamlss` (Stasinopoulos and Rigby, 2007). Inference in GAMLSS regression models is based on penalized maximum likelihood estimation, achieved via back-fitting loops over the additive predictor components. There are two algorithms available in the above software to maximize the penalized likelihood defined in (2.10): i) the `CG` algorithm, and ii) the `RS` algorithm. The first one, is a generalization of the algorithm proposed by Cole and Green (1992) and the last one comes from an extension of the Mean and Dispersion Additive Models' algorithm (MADAM, Rigby and Stasinopoulos, 1996). Note that sometimes the `RS` algorithm can be slow to converge but it is more adequate than `CG` algorithm when the parameters to estimate - in GAMLSS formulation the location, scale and shape parameters: $(\mu,\sigma,\upsilon,\tau)$ - are orthogonal (Stasinopoulos and Rigby, 2007). We refer to Rigby and Stasinopoulos (2005) for further details.

Although the GAMLSS algorithm (Rigby and Stasinopoulos, 2005; Stasinopoulos and Rigby, 2007) has several advantages (such as the modularity of the fitting procedure; the easy implementation of new response distributions and the facility to consider new additive terms); Klein et al. (2014b) has shown several disadvantages of this implementation. First, `gamlss`-package is relying on Newton-Raphson algorithm. The use of it implies that the negative second derivatives of the log-likelihood with respect to the predictors (i.e, $\tilde{w}_i = -\frac{\delta^2 l}{(\delta\eta_i)^2}$) are considered. However, Klein et al. (2014b) showed that the expected values of $w_i = \mathbb{E}(-\frac{\delta^2 l}{(\delta\eta_i)^2})$ are positive for several distributions; and this is not always true for $\tilde{w}_i$, such that several matrices involved might no longer be invertible (see supplementary material of Klein et al. (2014b) for the proof of the positive

definiteness). This fact can lead to numerical instabilities. This frequentist formulation could also present some problems when high dimensional effect terms are considered. The CGAMLSS formulated in Stasinopoulos and Rigby (2007) also provides confidence intervals for the estimates obtained based on the asymptotic normality assumptions of the maximum-likelihood estimator. However, in many situations these intervals are to narrow (Klein et al., 2015).

GAMLSS regression models are also contained as a special case in the `mgcv`-package (Wood, 2017) but only three possible response distributions are available. The current version of the `mgcv`-package only supports the following distributional models (Wood et al., 2016): i) *Gaussian location-scale model* where the mean and the standard deviation are both modelled using smooth linear predictors; ii) a *Generalized Extreme Value (GEV) model* where the location, scale and shape parameters are each modelled using a linear predictor; and finally iii) a *two-stage zero inflated Poisson model*.

2.3.2 Boosting inference

Mayr et al. (2012a) have proposed a boosting algorithm for high dimensional GAMLSS regression models which allows to deal with variable selection. This algorithm is available in the `gamboostLSS`-package. This approach solves some of the limitations of GAMLSS regression models and it could be useful in datasets with a large number of covariates. However, the main drawback of `gamboostLSS`-package is the computationally expensive cost. More details about this approach are shown in Chapter 5.

2.3.3 Bayesian inference: Structured additive distributional regression models

Distributional regression models can also be inferred using fully Bayesian (FB) methods, employing Markov Chain Monte Carlo (MCMC) simulation techniques (Klein et al., 2014a). Unknown variance or smoothing parameters are considered random variables with suitable hyperpriors, and estimated jointly with unknown functions and covariate effects using computationally efficient extensions of the MCMC techniques developed by Klein et al. (2014a).

Klein et al. (2014b) show that Bayesian inference based on MCMC outperforms the CGAMLSS frequentist approach (via simulation studies, Klein et al., 2014b): i) the predictors may include complex and hierarchical spatial effects and could be used to model hierarchical data situations. In this framework, MCMC works quite well, even in high-dimensional settings such as structured additive distribution regression models with a large number of unknown coefficients (Klein et al., 2014a). See Chapter 3 for an illustration example. ii) Furthermore, the Bayesian approach leads to reliable credible intervals - without relying on asymptotic arguments (Klein et al., 2015) - especially compared to situations in

which the asymptotic likelihood theory fails and the estimated points at least of similar quality (Klein et al., 2014b).

For all of these reasons, we will focus on structured additive distribution regression models. In the remainder of the thesis, we will refer to structured additive regression models (Klein et al., 2015) by DR in order to simplify the notation.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 3

Detecting differences in blood potassium concentrations by using a spatial distributional regression model

As mentioned in Chapter 1, some of the contents of this chapter were published in Espasandín-Domínguez et al. (2018a).

3.1 Introduction

Clinical laboratories contribute towards the screening, diagnosis and monitoring of many types of health condition. While it is believed that diagnostic testing may account for just 2% - 4% of all healthcare spending, it may influence 60% - 80% of medical decision-making (Hallworth, 2011).

Recently, general practitioners working in the *Santiago de Compostela Health Area* (SCHA) in northwestern Spain raised concerns over the high percentage of patients whose serum potassium concentrations were above the normal range, and over differences in the values recorded from one area to another. Analytical laboratories are commonly called upon to determine serum potassium concentrations, especially for patients with diabetes, heart and kidney disease. When potassium concentrations are recorded falsely as high (pseudohyperkalaemia) owing to specimen-collection or processing errors, medical mistakes can be made with disastrous consequences for patients. Although the list of sample management factors that can modify the potassium concentration is large, problems can be prevented by good laboratory practice (Stankovic and Smith, 2004).

The Clinical and Laboratory Standards Institute (2008) recommends that procedures be established for the transport of samples to laboratories to ensure that they are protected from deterioration (Tanner et al., 2008; Horowitz, 2008). The

extraction centres within SCHA are, however, up to 70 km away, and timely transport of samples to the laboratory is a challenge. The aim of the present work was to determine whether any geographical differences exist in terms of recorded serum potassium concentrations and their variability that might be attributed to pre-analytical factors, such as the centre where blood was extracted, adjusting for other potential covariates that might influence the results. For this purpose, a structured additive distributional regression model (see Chapter 2 and Klein et al., 2015) was used.

An advantage of this kind of model is the possibility of incorporating spatial effects. However, in most cases the output of spatial effects is not directly interpretable by biomedical researchers. This chapter proposes a way in which spatial effects can be visualised.

Another advantage is the possibility of contemplating a wide range of response variables. The deviance information criterion (DIC, Spiegelhalter et al., 2002; Klein et al., 2015) is commonly used for model choice in distributional regression. Quantile residuals can be used to check the performance of a selected model (Klein et al., 2015). In practice, the residuals can be assessed graphically in terms of quantile-quantile plots (Q-Q plots). However, interpreting the resulting graphs can be difficult, and the decision on the adequacy of a model remains subjective. Sometimes, even though the model is correct, the plot may deviate substantially from a straight line (Augustin et al., 2012). We therefore here propose the use of quantile-quantile plots with reference bands. To construct these bands, the methodology of Augustin et al. (2012) was adapted to the context of distributional regression by the first time.

In Section 3.2, we present the description of the database used in the study. Section 3.3 introduces the structured additive distributional regression models. Section 3.4 discuses the choice of the response distribution (and thus the model selected) and the construction of quantile-quantile plots with reference bands. Furthermore, a simulation study is presented in this section to asses the performance of this type of graphics. Finally, Section 3.5 discusses the results obtained following the analysis of the potassium database discussed in Section 3.2.

3.2 Data description

The database used in this chapter was provided by the Clinical Analysis Laboratory of the *Complejo Hospitalario Universitario de Santiago de Compostela* (CHUS). This supplied information on all blood extractions performed between 1 June and 31 December 2015 for which serum potassium, sodium and creatinine measurements were made. The initial number of samples was 145960. Those samples showing signs of haemolysis were excluded, as were those with creatinine or sodium concentrations outside the normal range (indicators of impaired kidney function). The final number of samples used in the present analysis was

therefore 95096 (see Figure 3.1).

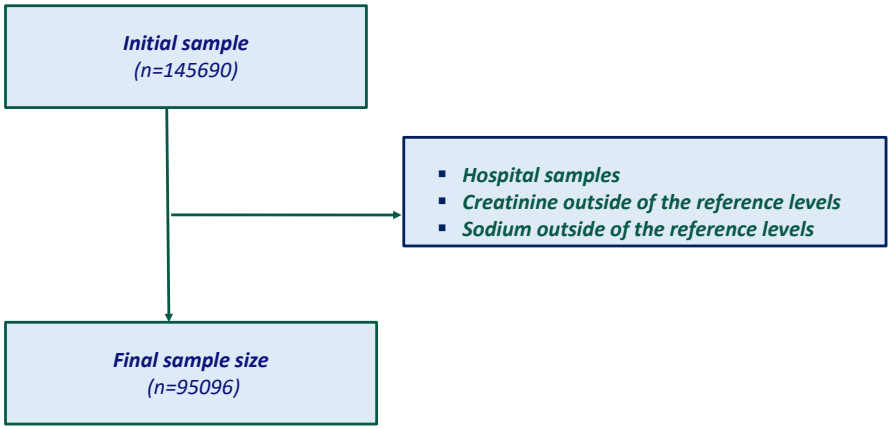


Figure 3.1: Application of the exclusion criteria.

The Health Area of Santiago de Compostela (SCHA), in Spain's northwest, covers 46 municipal districts (see Table 3.1 and Figure 3.2 for a map and the distribution of the population). The reference hospital is located in the city of Santiago de Compostela, from where the SCHA's health centres and doctors' practices are coordinated (see Figure 3.2 for their locations).

Blood samples were taken, usually daily, at designated locations in these 46 areas and transported by road to the CHUS following different routes.

The following variables were considered to be covariates: *gender*, *age* (in years), clot-contact time in minutes (*cctime*), and demographic information on the district where the extraction centres were located (s).

56% of the patients who provided blood samples were female, and 44% were male. The age range of the patients was 1-103 years. The mean (respectively SD) age was $54.8\ (19.0)$ years.

The clot-contact time was taken as the difference between the starting time of sample collection at the extraction centre (at 8.00 AM) and the entry time of the sample in the laboratory registry. The range was 4-458 min, the mean (SD) clot-contact time 229 (49) minutes, and the median time 231 minutes.

Patient serum potassium concentration - 2.1-7.1 mmol/L - will be understood as the response variable. The mean (SD) concentration was 4.6 (0.4) mmol/L, and the median 4.6 mmol/L.

The normal concentration of potassium (K) in the extracellular fluid is 3.5-5.3 mEq/L. Large deviations from these values are not compatible with life. Approximately 90% of the daily K intake is excreted in the urine, and a smaller percentage (10%) is excreted by the gastrointestinal tract. Therefore, within the body, the kidney is the major organ responsible for K homeostasis.

In these situations, hyperkalaemia needs to be distinguished from pseudohyperkalaemia. Pseudohyperkalaemia is the result of release of K from cells during the phlebotomy procedure, or specimen processing. The main causes of hyperkalaemia are cellular redistribution of potassium and impaired renal excretion. It is difficult to ingest enough K to become hyperkalaemic in the presence of normal renal and adrenal function. Dietary intake as a contributor to hyperkalaemia is common in the setting of impaired kidney function (Palmer and Clegg, 2016). As mentioned before, recordings with creatinine levels outside of the normal range were excluded.

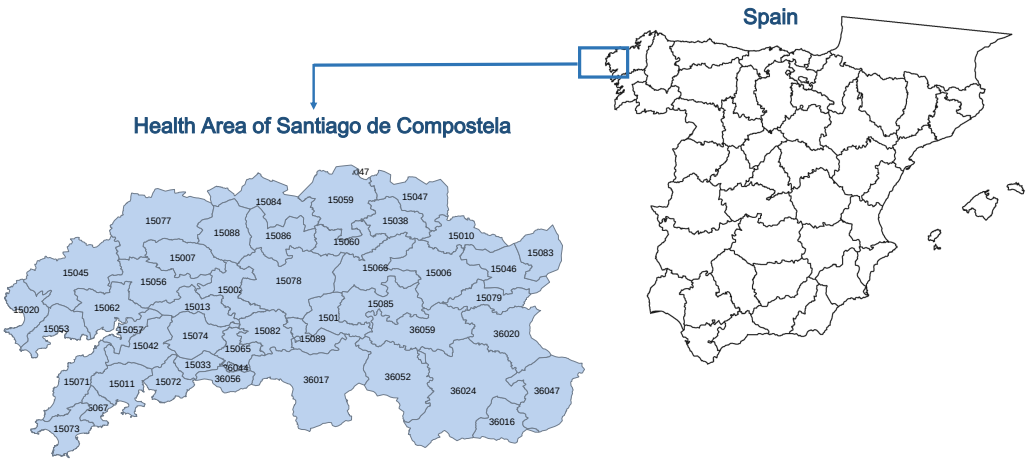


Figure 3.2: Health Area of Santiago de Compostela. Codes are in Table 3.1.

3.3 Structured additive distributional regression models

In this section, we present structured additive distributional regression type of models. More details and references can be found in Fahrmeir et al. (2013), Klein et al. (2014b) and Klein et al. (2015), among others.

Let us assume again that observations $(y_i, \boldsymbol{\nu}_i, i = 1, \ldots, n)$ are made, where y_i are observations on the response variable, and $\boldsymbol{\nu}_i$ represents the generic covariate vector. In this scenario, the response variables y_i can be assumed independently distributed with K-parametric densities $p(y_i \mid \vartheta_{i1}, \ldots, \vartheta_{iK}) \equiv p_i$. In other words, the conditional distribution p_i of an observation y_i given $\boldsymbol{\nu}_i$ is expressed in terms of the K distributional parameters of the response distribution: $\vartheta_{i1}, \ldots, \vartheta_{iK}$.

In structured additive distributional regression models, each parameter ϑ_k, $k = 1, \ldots, K$, of the response distribution is related to a semiparametric additive predictor $\eta_i^{\vartheta_k}$ defined in terms of the covariate vector $\boldsymbol{\nu}_i$. As in other types

Table 3.1: Districts in the *Santiago de Compostela Health Area* (SCHA), their codes, demographic characteristics, and the percentage of patients whose potassium results fell outside the normal range. The SCHA occupies an area of 4095 km^2; its population was 497171 at the time of the study (Instituto Galego de Estatística, 2015). The population density vary considerably; for example, the district of Santiago de Compostela, which has an area of 220 km^2, is home to 95612 people, while that of Toques, with an area of just 77.9 km^2, has 1213 inhabitants. The range of potassium values across the different districts is very wide, and not easily explained by the ageing of the population or differences in the prevalence of chronic diseases. *Source: www.ige.es*.

District	*Population*	*Area* (km^2)	Individuals out of range (%)	*District*	*Population*	*Area* (km^2)	Individuals out of range (%)
15002 - Ames	30267	80.0	18.8	15071 - Porto do Son	9436	94.6	24.3
15006 - Arzúa	6219	155.5	16.1	15072 - Rianxo	11386	58.8	21.2
15007 - A Baña	3698	82.3	22.4	15073 - Ribeira	27372	68.8	26.0
15010 - Boimorto	2125	86.6	17.8	15074 - Rois	4710	92.8	21.1
15011 - Boiro	18950	86.6	26.9	15077 - Santa Comba	9635	203.7	30.3
15012 - Boqueixón	4321	73.2	20.3	15078 - Santiago de Compostela	95612	220.0	14.9
15013 - Brión	7564	74.9	19.0	15079 - Santiso	1709	67.4	19.0
15020 - Carnota	4284	70.9	28.0	15082 - Teo	18505	79.3	12.6
15033 - Dodro	2882	36.1	20.8	15083 - Toques	1213	77.9	20.2
15038 - Frades	2460	81.6	28.1	15084 - Tordoia	3591	124.6	21.2
15042 - Lousame	3463	93.6	20.1	15085 - Touro	3778	115.3	15.5
15045 - Mazaricos	4173	187.3	23.1	15086 - Trazo	3263	101.3	32.0
15046 - Melide	7538	101.3	29.1	15088 - Val do Dubra	4033	108.6	29.1
15047 - Mesía	2734	107.1	31.5	15089 - Vedra	5059	52.8	22.7
15053 - Muros	8960	72.9	21.5	36016 - Dozón	1174	74.2	31.1
15056 - Negreira	6936	115.1	17.5	36017 - A Estrada	21025	280.8	18.6
15057 - Noia	14472	37.2	20.7	36020 - Agolada	2585	147.9	31.4
15059 - Ordes	12776	157.2	15.1	36024 - Lalín	20005	326.8	22.3
15060 - Oroso	7413	72.6	18.6	36044 - Pontecesures	3062	6.7	12.8
15062 - Outes	6691	99.7	21.2	36047 - Rodeiro	700	154.9	31.3
15065 - Padrón	8643	48.4	22.2	36052 - Silleda	8772	168.0	19.7
15066 - O Pino	4706	132.1	21.8	36056 - Valga	6062	40.6	20.1
15067 - Pobra	9623	32.5	20.4	36059 - Vila de Cruces	5556	155.0	21.9

of classic regression model, such as generalized linear regression models, a suitable response function is used to map the predictor to the parameter of interest, $\vartheta_{ik} = h^{\vartheta_k}(\eta_i^{\vartheta_k})$. According to Klein et al. (2015), the superscript ϑ_k refers to the fact that K predictors specific, for each of the distribution parameters of the response variable (and not just for the mean as in classical regression), are taken into account. Moreover, for an observation $i = 1, \ldots, n$, a suitable structured additive predictor for parameter ϑ_k can be written as

$$\eta_i^{\vartheta_k} = \beta_0^{\vartheta_k} + f_1^{\vartheta_k}(\boldsymbol{\nu}_i) + \cdots + f_{J_k}^{\vartheta_k}(\boldsymbol{\nu}_i) + f_{spat}^{\vartheta_k}(s_i), \tag{3.1}$$

where $\beta_0^{\vartheta_k}$ represents the overall level of the predictor, and the functions $f_j^{\vartheta_k}(\boldsymbol{\nu}_i)$, $j = 1, \ldots, J_k$ represent the different covariate effects of subsets of $\boldsymbol{\nu}_i$. Note that each distribution parameter may depend on different covariates and a different number of effects, say J_k. The generic representation with the complete covariate vector can be used to simplify the notation. Finally, $f_{spat}(s)$ is the spatial effect capturing heterogeneity at the level of the districts s.

In structured additive regression, each function f_j is approximated by a linear combination of D_j appropriate basis functions

$$f_j(\boldsymbol{\nu}_i) = \sum_{d_j=1}^{D_j} \beta_{j,d_j} B_{j,d_j}(\boldsymbol{\nu}_i).$$

In matrix notation, we can write $\boldsymbol{f}_j = (f_j(\boldsymbol{\nu}_1), \ldots, f_j(\boldsymbol{\nu}_n))' = \boldsymbol{Z}_j\boldsymbol{\beta}_j$ where $Z_j[i, d_j] = B_{j,d_j}(\boldsymbol{\nu}_i)$ is an $n \times D_j$ design matrix and $\boldsymbol{\beta}_j$ is the vector of coefficients (with dimension D_j) to be estimated. The basis function representation then lead us to the following matrix representation of the predictor (3.1)

$$\boldsymbol{\eta} = \beta_0\mathbf{1} + \boldsymbol{Z}_1\boldsymbol{\beta}_1 + \cdots + \boldsymbol{Z}_J\boldsymbol{\beta}_J. \tag{3.2}$$

For each of the parameter vectors $\boldsymbol{\beta}_j$, the multivariate normal prior can be assumed

$$p(\boldsymbol{\beta}_j \mid \tau_j^2) \propto \left(\frac{1}{\tau_j^2}\right)^{\frac{rk(\boldsymbol{K}_j)}{2}} \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j\right),$$

in which the (potentially rank-deficient) precision matrix $\boldsymbol{K}_j$ corresponds to the penalty matrix in a frequentist formulation. Note that here, we are using a generic notation for different terms (e.g penalized splines, Markov random fields, random effects). The precision matrix, $\boldsymbol{K}_j$, for the Markov random field is rank-deficient by construction since only deviations from a constant spatial effect are penalized. This leads to a rank deficiency of one. In the case of the penalized splines, the rank deficiency comes from the fact that polynomials of degree equal to the difference order minus one are not penalized. However for other possible terms like random effects this matrix is not rank-deficient.

A prior smoothing variance of τ_j^2 is assigned as an inverse gamma hyperprior $\tau_j^2 \sim IG(a_j, b_j)$ (with $a_j = b_j = 0.001$ as a default option in order to obtain data-driven smoothness). Small values are usually assumed for the hyperparameters a_j and b_j (to be close to Jeffreys' noninformative prior, Berger et al., 2009; Jeffreys, 1998). In some situations (such as small sample sizes), the estimation of models with different values for a_j and b_j are then recommend.

Again following Klein et al. (2015), and to simplify the notation, the dependence on the distributional parameter indicated by the superscript ϑ_k, the observation index i, and the function index j, have been dropped.

Fahrmeir et al. (2013) discuss all terms included in this generic predictor. The following paragraphs outline the prior assumptions for the hierarchical predictor required for potassium application. See Lang and Brezger (2004) for more details.

3.3.1 Linear effects and continuous covariates

For the effect of the intercept, β_0, and the gender of the individuals, β_1, a flat, non-informative prior was assumed.

The non-linear effects of continuous covariates (*age* and *clot-contact time*) were modelled using Bayesian versions of penalized splines (P-splines, Lang and Brezger, 2004), introduced into a frequentist setting by Eilers and Marx (1996). To model age and clot-contact time, 20 inner knots, a cubic spline basis, and a second order random walk prior for penalized splines were contemplated.

For the penalized spline specifications, we were able to rely on extensive research concerning the number and placement of knots, the order of the random walk prior, and the degree of the polynomial spline, e.g. Eilers and Marx (1996), Lang and Brezger (2004) and Brezger and Lang (2006). Their main findings can be summarized as follows: i) The number and placement of the knots has only a very minor impact on the fit if the number of knots chosen is not too small. ii) 20 equidistant knots yield sufficient flexibility for basically all situations of applied interest. iii) Second order random walk priors leave a linear effect unpenalized which is in analogy to the common penalty for smoothing splines. Moreover, first order differences often yield more wiggly estimates. iv) Finally, cubic splines yield a visually smooth function estimate which is twice continuously differentiable. This fits very well with the common visual perception of non-linear effects.

3.3.2 Spatial effects

In some biomedical studies may be useful to decompose spatial effects in two types: i) a conditional structured part, and a ii) non conditional unstructured part which allows to explain the heterogeneity among possible different regions. By estimating both structured and unstructured components, we can distinguish

between possible strong spatial structures and others patterns present only locally. An advantage of this kind of model is the possibility of incorporating both types of spatial effects. However, in most cases the output of spatial effects is not directly interpretable by biomedical researchers. This chapter explains a way to introduce spatial effects into distributional regression framework and proposes a way in which spatial effects can be visualised.

In this work, the spatial effects, f_{spat}, were understood as the sum of the spatially structured correlated (smooth) effects, f_{str}, and spatially uncorrelated (unsmooth) effects, f_{unstr}:

$$f_{spat}(s) = f_{str}(s) + f_{unstr}(s).$$

A spatial effect is usually a surrogate of many unobserved influential factors, some of which may obey a strong spatial structure while others may be present only locally. By estimating a structured and an unstructured component, it was hoped that distinctions could be made between these kinds of influential factor (Besag et al., 1991).

Structured spatial effects

For correlated spatial effects (or structured spatial effects), we assume spatial correlations defined implicitly by assuming a Markov random field (Fahrmeir et al., 2013) as a prior distribution for the separate regression coefficients corresponding to the distinct regions. The Markovian structure is determined by the neighbourhood structure for the regions and the precise form of the prior distribution is defined by:

$$\beta_{str,s}|\beta_{str,r}, r \neq s, \tau^2_{str} \sim N\left(\frac{1}{N_s}\sum_{r\in\delta_s}\beta_{str,r}, \frac{\tau^2_{str}}{N_s}\right),$$

where $N_s = |\delta_s|$ is the number of adjacent sites or neighbours, and $r \in \delta_s$ denotes that region r is a neighbour of site s. The conditional mean of $\beta_{str,s}$, given all other coefficients, is the average of the neighbouring regions.

The joint distribution of these spatial effects can be considered as follows

$$p(\boldsymbol{\beta} \mid \tau^2_{str}) \propto \left(\frac{1}{\tau^2_{str}}\right)^{\frac{d-1}{2}} \exp\left(-\frac{1}{2\tau^2_{str}}\boldsymbol{\beta}'\boldsymbol{K}\boldsymbol{\beta}\right),$$

where d denotes the total number of regions. The precision matrix $\boldsymbol{K}$ is a penalty matrix defined as

$$K[s,r] = \begin{cases} -1 & \text{when } s \neq r, s \sim r \\ 0 & \text{when } s \neq r, s \not\sim r \\ |N(s)| & \text{when } s = r. \end{cases}$$

In this definition, $\sim$ denotes that regions s and r are neighbours, and $\not\sim$ denotes that s and r are not neighbours.

For structured spatial effects, the design matrix $\boldsymbol{Z}$ of predictor (3.2) is an indicator matrix which links the observations with the corresponding districts as follows

$$Z[i,s] = \begin{cases} 1 & \text{if } y_i \text{ is observed in region } s \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

Unstructured spatial effects

Additional uncorrelated random effects (or unstructured spatial effects) may be incorporated as a surrogate for unobserved local small-area, group or individual specific heterogeneity. If one ignores spatial proximity and interprets $s \in \{1, \ldots, S\}$ as a cluster variable that only represents membership to different groups (such as different individuals in longitudinal data or more generally to different clusters of observations), we can assume a standard Gaussian random effects prior, i.e. $f_{unstr}(s) \sim N(0, \tau^2_{unstr}), s \in \{1, \ldots, S\}$, where the different groups correspond to the different administrative regions in the data set.

Note that here two variances are used for the structured, τ^2_{str}, and the unstructured effect, τ^2_{unstr}. Structurally, both variances are of the same type but they refer to different prior assumptions, where one assumes spatial structure (τ^2_{str}) while the other one assumes spatial independence (τ^2_{unstr}).

3.4 Inference and choice of the response distribution

Distributional regression models can be inferred employing computationally efficient extensions of the MCMC techniques developed by Klein et al. (2014a). See Chapter 2 and Klein et al. (2014a) for more details.

Many approaches have been proposed for dealing with model choice in a Bayesian framework. In the present chapter, the Deviance Information Criterion (DIC, Spiegelhalter et al., 2002; Klein et al., 2015) was used to choose the best response distribution. The DIC is a commonly used criterion for model choice in Bayesian inference. It became popular in part because of its easy implementation from the MCMC output. The performance of the DIC was valued as positive by (Klein et al., 2015), who compared several misspecified models with the true DIC model.

A rule of thumb says that DIC differences of 10 and more between two competing models indicate the model with the lower DIC to be superior (Klein et al., 2015). In the present work, the DIC showed the log-normal distribution to provide the best and most parsimonious fit (see Table 3.3).

Table 3.2: Selected candidate distributions response. The response function is usually chosen to ensure appropriate restrictions on the parameter space: *exponential function* to ensure positivity and *identity function* if the parameter space is unrestricted. In this table, $\Gamma(\sigma) = (\sigma - 1)!$ and Φ symbolizes the density function of a standard normal distribution.

Distributions	**Density**	**Parameters**	**Response functions**
Normal	$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$	$\mu \in \mathbb{R}$ $\sigma^2 > 0$	$h^{\mu}(\eta) = \eta$ $h^{\sigma^2}(\eta) = e^{\eta}$
Log-normal	$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2 y}} \exp\left(-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right)$	$\mu \in \mathbb{R}$ $\sigma^2 > 0$	$h^{\mu}(\eta) = \eta$ $h^{\sigma^2}(\eta) = e^{\eta}$
Truncated-normal	$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \frac{1}{\sigma\left(-\Phi(\frac{-\mu}{\sigma})\right)}$	$\mu \in \mathbb{R}$ $\sigma^2 > 0$	$h^{\mu}(\eta) = \eta$ $h^{\sigma^2}(\eta) = e^{\eta}$
Inverse Gaussian	$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2 y^3}} \exp\left(-\frac{(y-\mu)^2}{2y\mu^2\sigma^2}\right)$	$\mu > 0$ $\sigma^2 > 0$	$h^{\mu}(\eta) = e^{\eta}$ $h^{\sigma^2}(\eta) = e^{\eta}$
Gamma	$p(y \mid \mu, \sigma) = \left(\frac{\sigma}{\mu}\right)^{\sigma} \frac{y^{\sigma-1}}{\Gamma(\sigma)} \exp\left(-\frac{\sigma}{\mu} y\right)$	$\mu > 0$ $\sigma^2 > 0$	$h^{\mu}(\eta) = e^{\eta}$ $h^{\sigma^2}(\eta) = e^{\eta}$

Table 3.3: Comparison of DIC values for the candidate distributions.

Distribution	**DIC**
Normal	87816.70
Gamma	86880.60
Inverse Gaussian	86872.60
Log-normal	86821.40

3.4.1 Quantile residuals and quantile-quantile plots

In addition to the DIC, residuals can also be used to check the performance of a selected model after estimation. In this framework, we will consider quantile residuals as proposed in Klein et al. (2015). The term "quantile residual" is due to Dunn and Smyth (1996).

Let $F_i(\cdot \mid \hat{\boldsymbol{\vartheta}}_i)$ be the fitted cumulative distribution with estimate parameters, $\hat{\boldsymbol{\vartheta}}_i = \left(\hat{\vartheta}_{i1}, \ldots \hat{\vartheta}_{iK}\right)$, plugged in. The quantile residuals (Dunn and Smyth, 1996) are defined by, $\hat{r}_i = \Phi^{-1}(u_i)$, where Φ is the cumulative distribution function of a standard normal and $u_i = F(y_i \mid \hat{\boldsymbol{\vartheta}}_i)$. If y_i is a realisation of a discrete response variable, u_i is a random number from the uniform distribution on the interval $\left[F_i(y_i - 1 \mid \hat{\boldsymbol{\vartheta}}_i), F_i(y_i \mid \hat{\boldsymbol{\vartheta}}_i)\right]$. Specifically, the following two transformations define quantile residuals. First, the estimated cumulative distribution function implied by the model is used to transform the observations into approximately independent uniformly distributed random variables. Second, the inverse of the

cumulative distribution function of the standard normal distribution is used to get variables which are approximately independent with standard normal distribution. These results assume that the model is correctly specified and parameters are consistently estimated. If not, quantile residuals are expected to exhibit detectable departures from the characteristic properties described above.

If the estimated model is close to the true model, the quantile residuals approximately follow a standard normal distribution, even if the model distribution itself is not a normal distribution (Dunn and Smyth, 1996). In practice, the residuals can be assessed graphically in terms of quantile-quantile plots. However, conclusions taken out from plots leave some room for subjectivity. In this particular setting, to consider p-values or tests for normality are not an appropriate way of making an objective decision in favour of or against normality. First of all, a test with normality as the null hypothesis will not allow us to prove normality but might at best give us an indication for the absence of evidence against normality. Secondly, we apply our model to a large data set such that even slight deviations from normality will turn out to be significant. Finally, p-values are also not very common in the Bayesian paradigm that we adopt in this chapter. As a consequence, we did not consider p-values and/or tests for normality. To improve the interpretability of this type of graphics, we propose to add reference bands around the diagonal line to give a rough indication of the uncertainty implied by estimating the model from finite data. More precisely, following Augustin et al. (2012)[1], we repeatedly simulate data from the fitted model and add pointwise minima and maxima for the resulting quantile residuals from a prespecified number of replications. However, from a pragmatic perspective it can be sufficient to demand that most of the points are in the reference bands. See Appendix A for more details about the construction of the reference bands.

Simulation study

In this section, we carry out a simulation study in R, (R Core Team, 2017), to check the performance of quantile-quantile plots (Q-Q plots) with reference bands for determining the goodness of fit in the framework of distributional regression, previously introduced in the above section. Once the quantile-quantile plots with reference bands were constructed, we have calculated the percentage of points that lies outside of the reference bands.

Below, we specify the scenarios that we considered in the simulation study. We focus on smooth effects of a single covariate on all involved predictors for the normal (N), log-normal (LN), skew normal (SN), gamma (GA), and inverse Gaussian (IG) distributions. Note that skew-normal is a three parameter distribution that allows for modelling skewness in addition to location and scale.

[1]The work of Augustin et al. (2012) proposes two alternative methods for generating quantiles for Q-Q plots, in this work we have follow the method based on direct simulation since our objective is to obtain reference bands.

Figure 3.3 plots the $SN(0, 1, \upsilon)$ distribution for different values of $\upsilon = 1/45, 0, 6$. Changing υ reflects the distributions about the origin, changing the skewness from positive to negative.

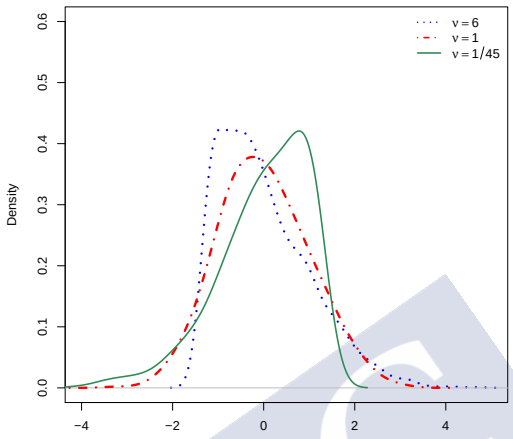


Figure 3.3: The skew normal distribution, SN(0, 1, υ), for different values of υ.

- We consider one continuous covariate obtained with $\nu_i \sim U[1, 5]$ for $i = 1, \ldots, n$ and sample sizes $n = 250, 500$, and 2000.
- Below we specify the response distribution of the six models considered in our simulation study:

  Normal scenario

  - *(M1)*: $y_i \sim N\left(\mu_i = f_1^{\mu}(\nu_i), \sigma_i = \exp\left\{f_2^{\sigma}(\nu_i)\right\}\right)$
  - *(M2)*: $y_i \sim SN\left(\mu_i = f_1^{\mu}(\nu_i), \sigma_i = \exp\left\{f_2^{\sigma}(\nu_i)\right\}, \upsilon = 1/45\right)$
  - *(M3)*: $y_i \sim SN\left(\mu_i = f_1^{\mu}(\nu_i), \sigma_i = \exp\left\{f_2^{\sigma}(\nu_i)\right\}, \upsilon = 6\right),$

  with $f_1^{\mu}(\nu) = 5 + (\nu - 3)^2$ and $f_2^{\sigma}(\nu) = 0.05\nu^2$.

  Log-normal scenario

  - *(M4)*: $y_i \sim LN\left(\mu_i = \exp\left\{f_1^{\mu}(\nu_i)\right\}, \sigma_i = \exp\left\{f_2^{\sigma}(\nu_i)\right\}\right)$
  - *(M5)*: $y_i \sim GA\left(\mu_i = \exp\left\{f_1^{\mu}(\nu_i)\right\}, \sigma_i = \exp\left\{f_2^{\sigma}(\nu_i)\right\}\right)$
  - *(M6)*: $y_i \sim IG\left(\mu_i = \exp\left\{f_1^{\mu}(\nu_i)\right\}, \sigma_i = \exp\left\{f_2^{\sigma}(\nu_i)\right\}\right),$

  with $f_1^{\mu}(\nu) = 5 + (\nu - 3)^2$ and $f_2^{\sigma}(\nu) = 0.05\nu^2$.

- We consider the following settings

- The predictors for each model are given by $\eta^{\mu} = \beta_0 + f_1^{\mu}(\nu)$ and $\eta^{\sigma} = \beta_0 + f_2^{\sigma}(\nu)$.
- In the normal scenario (models of type *M1, M2* and *M3*), the normal distribution has been assumed for the response distribution.
- In the log-normal scenario (models of type *M4, M5* and *M6*), the log-normal distribution has been assumed for the response distribution.
- For each scenario and each model, we consider sample sizes $n = 250$, 500, and 2000.
- The number of iterations steps of MCMC for each simulation run, $r = 1, \ldots, 1000$ is set to 12000 with burn-in phase of 2000 iterations. We store and use every 10-th iterate.
- Functions f_1 and f_2 have been estimated using cubic P-splines on a grid of 20 equidistant knots and second order random walk prior. We considered inverse gamma priors with parameters $a = b = 0.001$ for the smoothing variances τ^2 as the default choice for MCMC.
- The computations have been carried out in BayesX (Belitz et al., 2015) software.

The results are compared in terms of the percentage of points outside of the reference bands. In the following, we summarize the results obtained:

Normal scenario

(i) If the true distribution is a normal distribution and a normal model is estimated (*M1*); in more than 99.99% of replications all points are inside the reference bands for all the sample sizes considered ($n = 250, 500, 2000$). Figure 3.4 shows a Q-Q plot with 95% reference bands for a representative replicate ($n = 250, 500, 2000$), illustrating that Q-Q plots work very well in this framework.

(ii) Figures 3.5 and 3.6 summarize the results obtained if the true model is a skew normal with left skewness (models of type *M2*) or right skewness (models of type *M3*) and a normal distribution is estimated, in both cases. As we can see in these figures, the higher sample size, the easier it gets for Q-Q plots with reference bands to decide for the true distribution; resulting in more than 70% of points (for $n = 2000$) being outside of the reference bands in both, *M2* and *M3* type of models (see Table 3.4). Figure 3.6 shows a Q-Q plot with 95% reference bands for a representative replicate (and sample sizes $n = 250, 500, 2000$). This figure illustrates that the plots can detect the model misspecification.

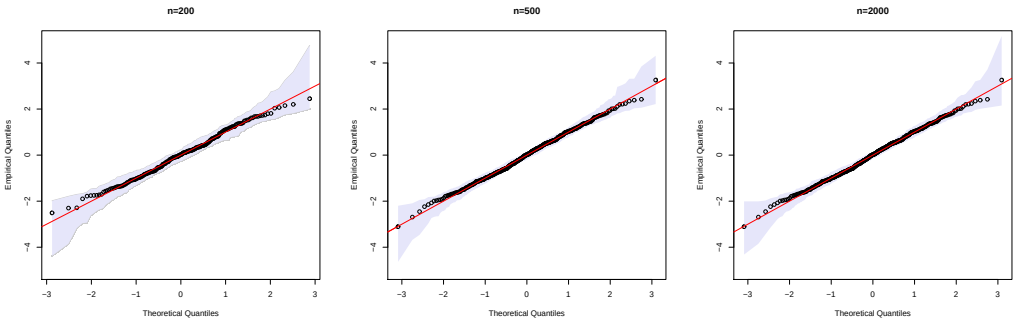


Figure 3.4: Quantile-quantile plots with 95% reference bands of quantile residuals for a representative replicate of a model of type *M1* and sample sizes, $n = 250, 500,$ and 2000. This figure shows that all points are in the reference bands.

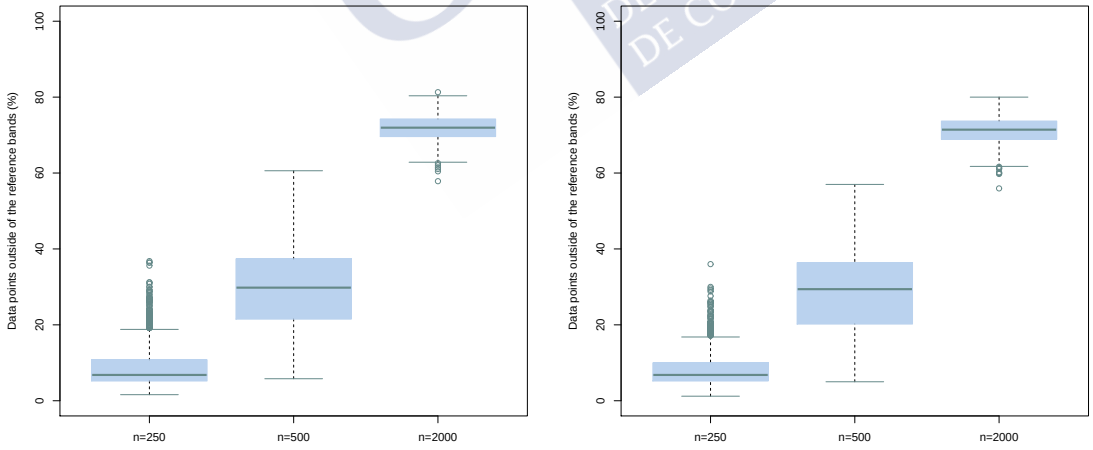


Figure 3.5: Comparison of percentage of points outside of the reference bands for the models of type *M2* (left) and *M3* (right) within the normal scenario and sample sizes, $n = 250, 500,$ and 2000.

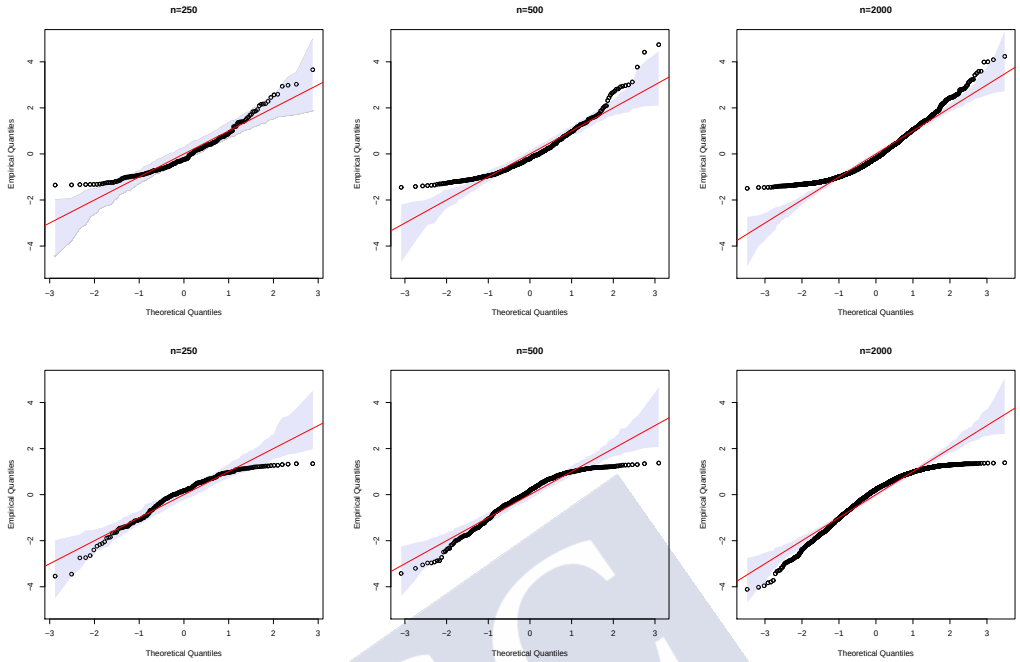


Figure 3.6: Quantile-quantile plots with 95% reference bands of quantile residuals for a representative replicate of a model of type *M2* (top) and *M3* (bottom) for sample sizes, $n = 250, n = 500, n = 2000$.

Log-Normal scenario

(i) If the true distribution is a log-normal distribution and it is also assumed a log-normal distribution in the estimated model (*M4*); in more than 99.99% of replications all points are inside the reference bands for $n = 250, 500, 2000$. (See Table 3.4 and Figure 3.7).

(ii) The results obtained fitting a gamma by a log-normal distribution (*models of type M5*) are summarized in Figure 3.9. As can be seen in Figure 3.10, the number of points in the tails increase with sample size, which indicates that Q-Q plots' good performance increase with sample size.

General conclusions

- Table 3.4 summarizes the results obtained in the simulations in terms of the percentage of data points outside of the reference bands.
- Q-Q plots with reference bands, clearly favours the true model in our simulation.
- The proposed Q-Q plots with reference bands could detect model misspecification.

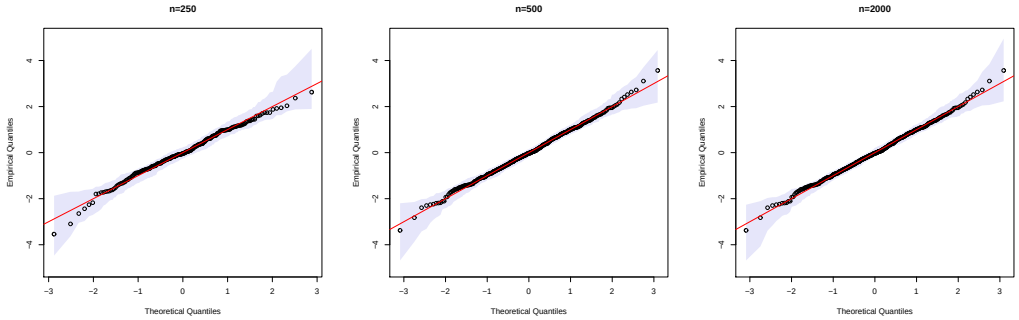


Figure 3.7: Quantile-quantile plots with 95% reference bands of quantile residuals for a replicate representative of a model of type *M4* for sample sizes, $n = 250, 500,$ and 2000.

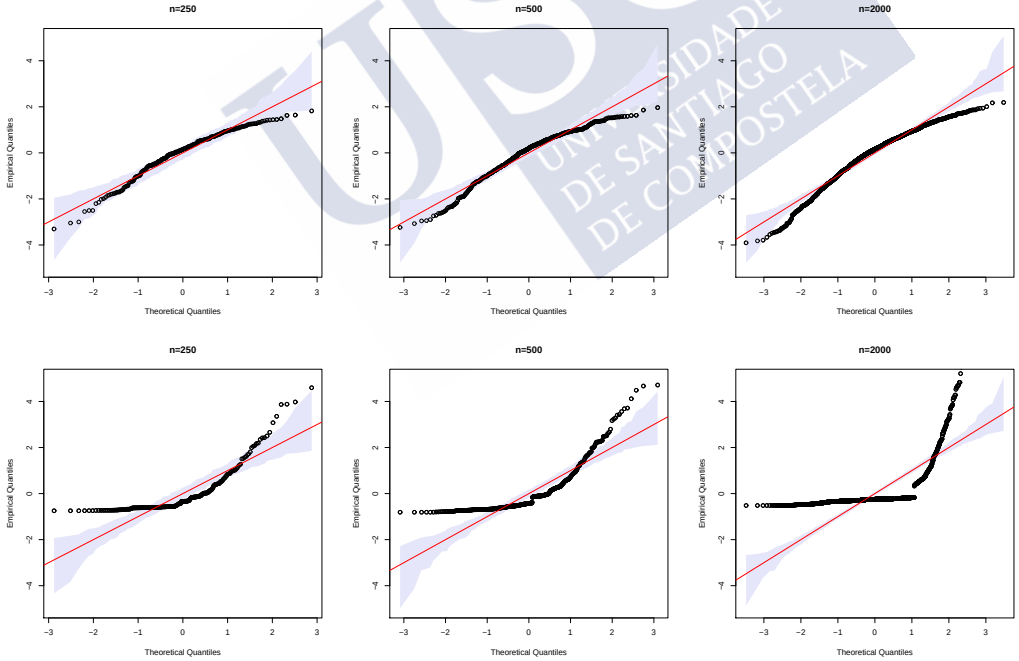


Figure 3.8: Quantile-quantile plots with 95% reference bands for a replicate representative and sample sizes, $n = 250, 500,$ and 2000, fitting a gamma (top) or a inverse-Gaussian (bottom) using a log-normal distribution illustrating that the proposed Q-Q plots can detect the model misspecification.

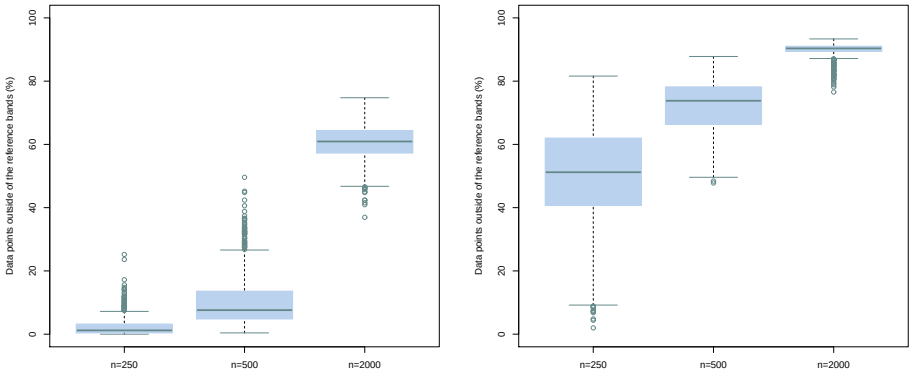


Figure 3.9: Comparison of percentage of points outside of the reference bands for models of type *M5* (left) and *M6* (right), on the log-normal scenario for sample sizes, $n = 250, 500,$ and 2000.

Table 3.4: Summary of the results obtained in the simulation study. Percentage of data points outside of the reference bands. In this table, SD denotes the standard deviation.

Framework		Minimum	Median	Mean	SD	Maximum
M1	$n = 250$	0.000	0.000	0.002	0.037	0.800
	$n = 500$	0.000	0.000	0.001	0.014	0.200
	$n = 2000$	0.000	0.000	0.002	0.026	0.600
M2	$n = 250$	1.200	6.800	8.388	4.938	36.000
	$n = 500$	5.000	29.400	28.355	10.749	57.000
	$n = 2000$	55.950	71.425	71.238	3.605	80.000
M3	$n = 250$	1.600	6.800	9.292	6.042	36.800
	$n = 500$	5.800	29.800	29.329	10.787	60.600
	$n = 2000$	57.850	71.950	71.797	3.392	81.300
M4	$n = 250$	0.000	0.000	0.000	0.000	0.000
	$n = 500$	0.000	0.000	0.001	0.014	0.200
	$n = 2000$	0.000	0.000	0.003	0.024	0.400
M5	$n = 250$	0.000	1.200	2.210	2.734	25.200
	$n = 500$	0.400	7.600	10.254	7.860	49.600
	$n = 2000$	36.950	60.925	60.546	5.363	74.750
M6	$n = 250$	2.000	51.200	49.954	15.533	81.600
	$n = 500$	47.800	73.800	72.129	7.670	87.800
	$n = 2000$	76.550	90.350	90.002	1.929	93.350

- The performance of the proposed Q-Q plot to check the goodness of fit of a selected model is better at higher sample sizes.

3.5 Data analysis

Q-Q plots of quantile residuals with reference bands was positive valued in Section 3.4 to check the performance of a DR model.

Figure 3.10 shows the Q-Q plot for the selected model with a log-normal distribution. Figure 3.10 also provides reference bands for judging the relevance of departures of quantile-quantile plots from the ideal red line. The log-normal distribution turns out to be appropriate for residuals in the range between -2.5 and 2.7 but deviates from the diagonal line for extreme values. Note, however, that these extreme values correspond to only 1.37% of the total data (0.87% of the database to the left and 0.5% to the right) such that our model explains the vast majority of observations well.

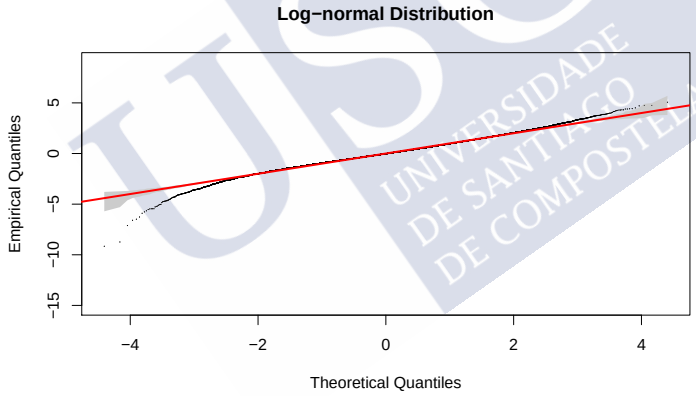


Figure 3.10: Quantile-quantile residuals plot for the selected model with reference bands: the closer the residuals to the bisecting red line, the better the fit to the data.

Thus, the structured additive distributional regression model described in Section 3.3 was used with a log-normal distribution response and with two covariate dependent parameters (corresponding to the mean of the log-transformed potassium concentrations and the scale parameter σ^2), to describe and compare the potassium concentrations recorded in the different districts of the SCHA. This model is expressed as follows:

$$\begin{cases} \eta^{\mu} = \beta_0^{\mu} + gender'\beta_1^{\mu} + f_1^{\mu}(age) + f_2^{\mu}(cctime) + f_{spat}^{\mu}(s) \\ \eta^{\sigma^2} = \beta_0^{\sigma^2} + gender'\beta_1^{\sigma^2} + f_1^{\sigma^2}(age) + f_2^{\sigma^2}(cctime) + f_{spat}^{\sigma^2}(s), \end{cases} \tag{3.4}$$

where β_0 represents the overall level of the predictor and β_1 captures the effect of the gender. Moreover, $f_1(age)$ and $f_2(cctime)$ are non linear effects of the age and the clot-contact time, respectively. Finally, $f_{spat}(s)$ is the spatial effect of the districts s.

3.5.1 Results

Statistical analyses were performed using open-source BayesX software (Belitz et al., 2015). The `BayesX` (Umlauf et al., 2018) and `R2BayesX` (Umlauf et al., 2015; Belitz et al., 2016) R-packages were used as graphic interfaces.

Table 3.5: Summary of estimated linear effects for model (3.4).

Parameter	mean	2.5% quantile	median	97.5% quantile
β_0^{μ} (intercept)	1.489	1.465	1.484	1.512
$\beta_0^{\sigma^2}$ (intercept)	-0.032	-0.050	-0.031	-0.014
β_1^{μ} (gender)	-0.006	-0.008	-0.006	-0.005
$\beta_1^{\sigma^2}$ (gender)	-0.032	-0.050	-0.032	-0.014

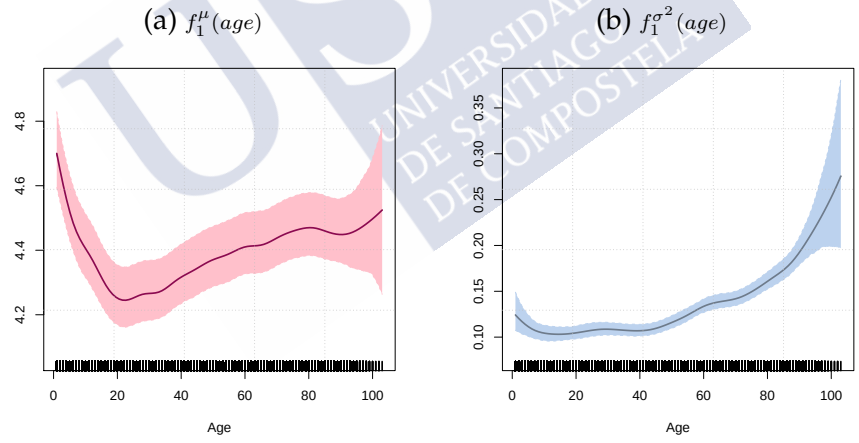


Figure 3.11: Posterior mean estimates of non linear effects of *age* on μ and σ^2.

All results are summarised in Table 3.5 and Figures 3.11-3.14. For the continuous variables, all covariates but the one that is visualised are fixed at their average while spatial effects are set to zero. For the spatial effects, we used district-specific averages for all covariates and determined significance based on the comparison with the average of all spatial effects. See Appendix A for more details.

Gender had some influence on the results (both on the mean and variance of the potassium concentrations), but this effect is clinically not relevant since the differences between both men and women were minimal (0.04 mg/dL). The

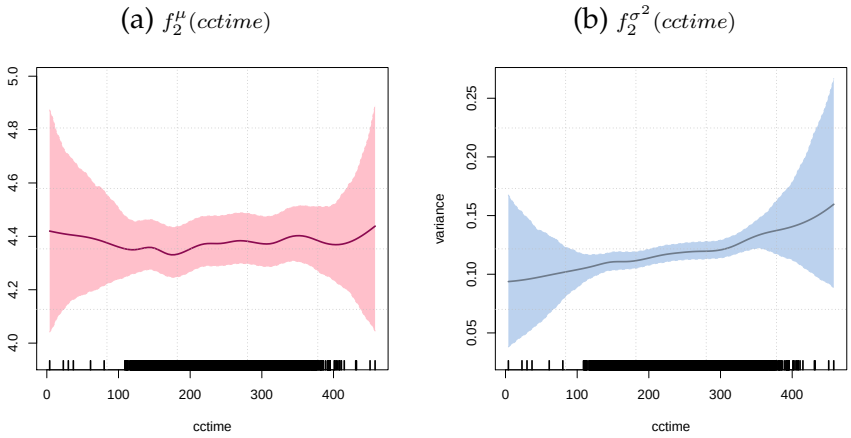


Figure 3.12: Posterior mean estimates of non linear effects of *cctime* on μ and σ^2.

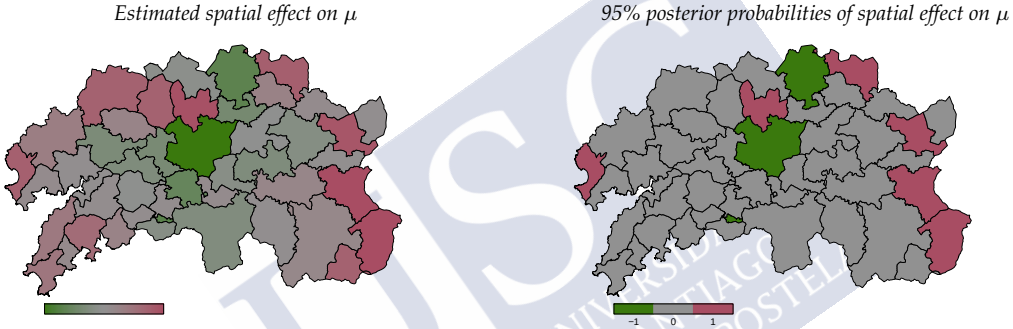


Figure 3.13: Posterior mean estimates of the complete spatial effects on mean potassium levels, f_{spat}^{μ}, and 95% posterior probabilities (right). In the right panel a value of 1 corresponds to a strictly positive 95% credible interval, and a value of -1 to a strictly negative credible interval; a value of 0 indicates that 0 is contained in the corresponding credible interval.

same for the variance. However, age, clot-contact time and the place of origin of the samples did influence the potassium concentrations recorded. Children and the elderly had higher mean potassium concentrations than did patients of intermediate age; the values recorded for the elderly also showed greater variability (Figure 3.11). These age-related findings might, however, be expected since venopunture is harder to perform in both children and the elderly, and both age groups show greater capillary fragility. This can lead to situations in which haemolysis occurs, releasing potassium from the red blood cells and increasing the recorded concentration for both age groups, as well as the variability of values recorded for the elderly.

The potassium concentrations recorded were clearly not uniformly distributed over the study area. In general, higher potassium concentrations were recorded in the areas farthest away from the test laboratory, although some areas close to

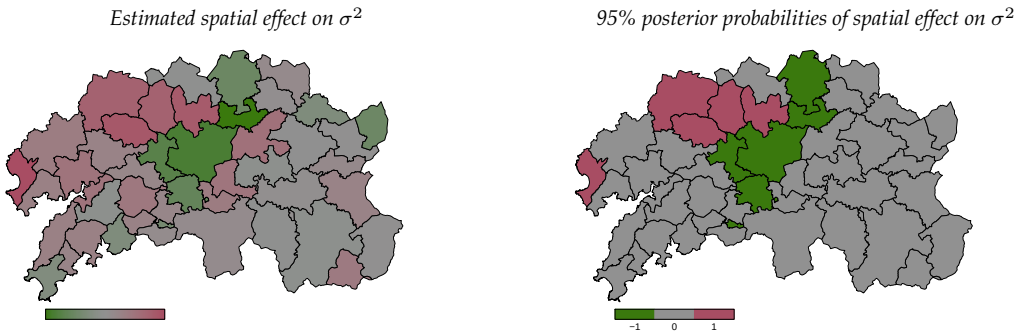


Figure 3.14: Posterior mean estimates of the complete spatial effects on the variance of potassium levels, $f^{\sigma^2}_{spat}$, and 95% posterior probabilities (right). In the right panel a value of 1 corresponds to a strictly positive 95% credible interval, and a value of -1 to a strictly negative credible interval; a value of 0 indicates that 0 is contained in the corresponding credible interval.

it also returned high values. In some areas, these high concentrations were accompanied by greater variability in the results, particularly in the districts on the northern periphery of the study area. The districts to the south-east also returned high potassium concentrations but with the less variability. Although caution should be exercised when interpreting these spatial analysis results, it may be that potassium concentrations in the periphery are related to pre-analytical factors associated with the extraction centres. The affected areas are also those with the lowest population densities; they are therefore likely to have less equipment, fewer personnel, and perhaps less well trained personnel than in the more central districts. These periphery districts may also have older inhabitants, clinical practices may be less homogeneous, and they are the worst communicated with the test laboratory (Figures 3.13 and 3.14).

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 4

Extensions to bivariate responses: Copula regression models

Diabetes is one of the most common human disorders. Early diagnosis and strict glucose control are crucial if serious complications are to be prevented or delayed. Prognoses for diabetes are based largely on determining the plasma glucose and glycated haemoglobin (HbA1c) concentrations. These tests are used to detect individuals with pre-diabetes, as well as to screen for and diagnose the disease. The results, however, are not foolproof, and the clinical usefulness of these tests is affected by a number of biological and analytical factors. In clinical practice, the introduction of other measures of glucose homeostasis, such as plasma fructosamine and glycated albumin, is attractive, especially when dealing with patients in whom the measurement of HbA1c may be biased (e.g., patients with kidney disease, anaemia, or disorders involving abnormal haemoglobin metabolism) (American Diabetes Association, 2018). Unfortunately, discordances are often seen among the results for HbA1c and other glycated proteins, and clinicians need to be aware of the conditions that might explain them (Sacks, 2011). Several authors have proposed metrics for quantifying the discrepancies between HbA1c and blood glucose in the form of glycation "gaps" or "indices" the difference between the measured HbA1c and that which would be predicted from another measure of glycaemic control using a linear regression model (Cohen et al., 2003). However, both the glycation gap and index values correlate strongly with the concentration of HbA1c, and require that the distribution of this concentration be assumed Gaussian.

The biomedical aims of the present chapter are: i) to identify variables that might affect the mean concentrations of HbA1c and fructosamine, and which influence the variation in their concentrations, and ii) to identify the factors that may cause discordance between results for the concentrations of these glycated proteins. It is hoped that this will help improve the diagnosis and treatment of diabetes.

Such aims require the investigation of statistical methods able to flexibly and

simultaneously examine the mean concentrations of the above proteins, their variability, and the relationship between them. This chapter reviews the available flexible regression models that meet these requirements. More specifically, in this chapter, bivariate copula generalized additive models for location, scale and shape (CGAMLSS) are presented, based on either frequentist (Marra and Radice, 2017a) or Bayesian (Klein and Kneib, 2016b) inference principles. These types of model extend univariate generalized additive models for location, scale and shape (GAMLSS, Rigby and Stasinopoulos, 2005) as well as univariate distributional regression, (Klein et al., 2015) to the field of multivariate responses. More specifically, CGAMLSS estimates the joint multivariate distribution of a response vector where each parameter characterising the joint distribution is modelled simultaneously and is conditioned by covariates. The multivariate distribution is constructed from different copulas that allow for different dependence structures (Sklar, 1959). CGAMLSS enables the modelling of all distributional parameters using additive predictors that encompass several types of covariate effect, such as the non-linear effects of continuous covariates, random effects, and interactions.

The statistical aims of this chapter are i) to review existing copula regression models and to compare them in terms of a simulation study and a real biomedical application; ii) to show the clinical usefulness of this type of models, and provides a basis for its interpretation.

4.1 Dependence modelling with copulas

Other conditional copula regression techniques using copula functions have been proposed as alternatives to CGAMLSS, but they only provide some of the latter's flexibility - either because they only allow for the consideration of normal marginals, (Sabeti et al., 2014) or because they fail to consider additive predictors (Acar et al., 2013). In the frequentist setting, attention must also be drawn to vector generalized additive models (VGAM) as an alternative to CGAMLSS (Yee and Wild, 1996). The estimation of VGAM models is carried out by fitting a vector additive model at each iteration of the Iteratively Re-weighted Least Squares (IRLS) algorithm (see for example, Yee and Wild, 1996, for more details). VGAMs permits each parameter of a bivariate non-standard response to be estimated in a flexible manner using an additive predictor. Yee (2015) proposes the use of copulas as a special class of bivariate distributions. However, smoothing parameter selection is more difficult and thus leading to a potentially greater bias than in CGAMLSS estimations in terms of the accuracy of the estimated copula parameter (Klein and Kneib, 2016b). In addition to VGAM, Vatter and Chavez proposed a two-stage approach in which the parameters of the marginal distributions and the copula are determined separately (Vatter and Chavez-Demoulin, 2015). In contrast, in CGAMLSS regression models, all parameters

are estimated simultaneously. Via simulation studies, Marra and Radice (2017a) showed that CGAMLSS is slightly more efficient than a two step estimator. A major advantage of the CGAMLSS algorithms proposed by Marra and Radice (2017a) and Klein and Kneib (2016b) is that they were created in a modular fashion, and therefore new parametric continuous marginal distributions and copula functions can be easily included. Another is that copula regression parameters are integrated into the estimation of the model coefficients to allow for more flexible dependence modelling. Further, the smoothing parameters are selected automatically.

Despite the flexibility of the CGAMLSS framework, the lack of interpretability of the distributional parameters of the response variables reduces the use of this kind of modelling in the clinical setting. Indeed, clinicians prefer to use a generalized additive model requiring the assumption of a Gaussian response because the results are easier to understand. Further, with most additive models, the effects of continuous covariates on the results are centred (Hastie and Tibshirani, 1990; Wood, 2006). As mentioned above, several papers have been published describing the statistical methodology of CGAMLSS in detail, including simulation studies and providing some example analysis (as per, Marra and Radice, 2017a; Klein and Kneib, 2016b; Radice et al., 2016), however there are very few manuscripts on real biomedical data. For this reason, in this thesis we will give guidance for clinical researchers on how to apply these type of regression models. The present work also highlights a way to visualize and interpret the results obtained with the novel regression models used in practice. Furthermore, there are no studies in the literature comparing CGAMLSS frequentist and Bayesian approach.

4.2 Bivariate copula regression models

CGAMLSS (Marra and Radice, 2017a; Klein and Kneib, 2016b) model the joint distribution of a pair of response variables (y_1, y_2) given covariates based on a copula specification for the dependence structure between the two responses. Given the nature of the problem presented in the introduction, this work focuses on the use of CGAMLSS with the pair of continuous random variables, y_1 and y_2 as the response variables. In Section 4.4, we will denote, $y_1 = HbA1c$ and $y_2 = fructosamine$.

In the CGAMLSS approach, the joint cumulative distribution function (cdf) of y_1 and y_2, given the covariate information - collected in $\boldsymbol{\nu}$ -, is expressed in terms of the marginal cdfs and a copula function C that binds them together. But, What are copulas? Following Nelsen (2006), copula functions can be interpreted from two points of view: "From one point a view, copulas are functions that join or 'couple' multivariate distribution functions to their one dimensional marginal distribution functions. Alternatively, copulas are multivariate distribution func-

tions whose one-dimensional margins are uniform on the interval (0,1)."

More specifically, let us assume a function, $C : [0,1]^2 \rightarrow [0,1]$ such that its domain is the unit square and its image is $[0,1]$. C will be a copula function if it verifies the following properties (Palaro and Hotta, 2006):

(i) $C(y_1, y_2) = 0,\ \forall (y_1, y_2) \in [0,1]^2 \Leftrightarrow y_1 = 0$ and/or $y_2 = 0$.

(ii) $C(y_1, 1) = C(1, y_2) = 1,\ \forall (y_1, y_2) \in [0,1]^2$.

(iii) $\forall (a_1, a_2), (b_1, b_2) \in [0,1]^2$, it is verifies that:

$$C(a_2, b_2) - C(a_1, b_2) - C(a_2, b_1) + C(a_1, b_1) \geq 0.$$

One immediate consequence of the previous definition is that a bivariate cumulative distribution function defined on the unit square and with uniform marginal distribution is a copula function.

More, specifically, Sklar's theorem guarantees that we can write

$$F(y_1, y_2 \mid \boldsymbol{\nu}) = C(F_1(y_1 \mid \boldsymbol{\nu}), F_2(y_2 \mid \boldsymbol{\nu}), \rho), \tag{4.1}$$

where $F_1(y_1 \mid \boldsymbol{\nu})$ and $F_2(y_2 \mid \boldsymbol{\nu})$ are the marginal cdfs of $y_1 \mid \boldsymbol{\nu}$ and $y_2 \mid \boldsymbol{\nu}$ which take values of $(0,1)$. $C(\cdot, \cdot \mid \boldsymbol{\nu})$ is a uniquely defined two-place copula function that does contain information about the association between the two outcomes solely and ρ is an association copula parameter measuring the dependence between the responses. Note, that $F_1(y_1 \mid \boldsymbol{\nu})$ and $F_2(y_2 \mid \boldsymbol{\nu})$ are uniformly distributed on $[0,1]$.

The different parametric copula functions proposed in the literature allow different types of dependence structure between the response variables (see Figure 4.1 for a graphical illustration). For example, the Clayton copula allows one to consider asymmetric structures of dependence when two random variables show a stronger positive association at smaller values than at larger values. The Joe or Gumbel copula, in contrast, addresses the opposite situation, in which two random variables with positive dependence show a stronger association at higher values. Rotated versions of the Clayton, Gumbel and Joe copulas also exist for modelling negative structures of dependence. Gaussian and Frank copulas, do not have any tail dependence.

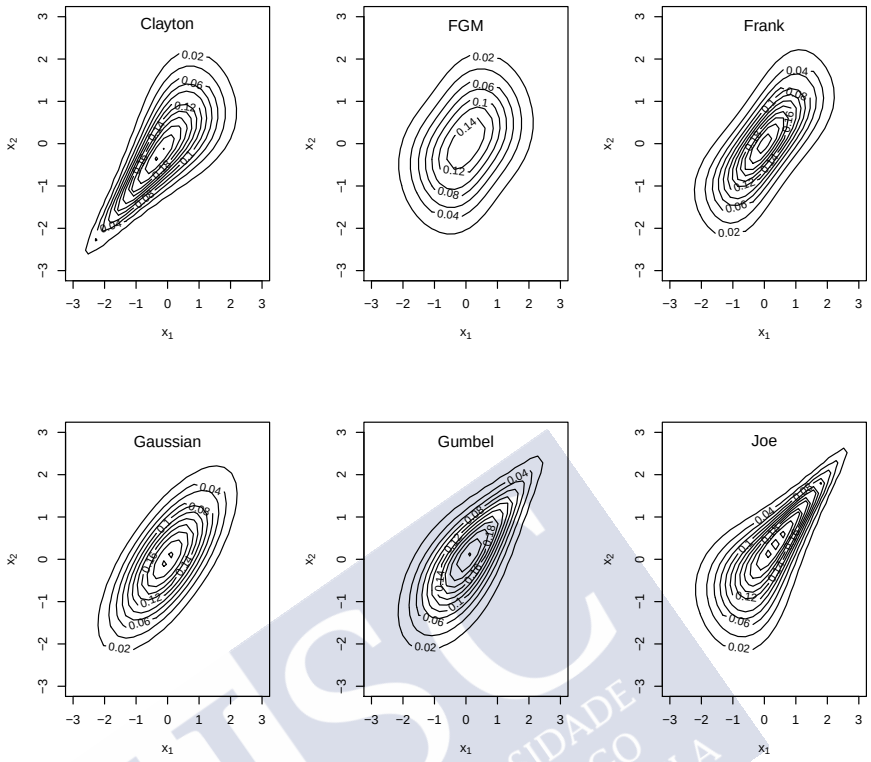


Figure 4.1: Contour plots of copula functions with standard normal margins for data simulated using a Kendall's τ coefficient of 0.5.

4.2.1 Model formulation

CGAMLSS regression models combine flexibility in the specification of the marginal distributions of a bivariate response vector with additional flexibility in the dependence structure induced by a copula. Further, by modelling each parameter of the response at the same time - and not just the marginal means - they allow for the quantification of regression effects on basically all aspects of the bivariate response distribution, including the location, scale, and the shape parameters of the marginal distributions, as well as on the copula parameter.

Let us assume that observations on the bivariate response vector $\{\boldsymbol{y}_i = (y_{i1}, y_{i2}), i = 1, \ldots, n\}$, where y_{i1}, y_{i2} are the marginal response variables, and the generic covariate vector $\{\boldsymbol{\nu}_i, i = 1, \ldots, n\}$ are available for n observational units. Let be p_1 and p_2 the marginal densities of y_1 and y_2, respectively,

$$p_{1,i} \equiv p_1\left(Y_{i1} \mid \left(\vartheta_{i1}^{(1)}, \ldots, \vartheta_{iK_1}^{(1)}\right)\right)$$
$$p_{2,i} \equiv p_2\left(Y_{i2} \mid \left(\vartheta_{i1}^{(2)}, \ldots, \vartheta_{iK_2}^{(2)}\right)\right).$$

Table 4.1: Some classic copulae, with corresponding parameter range of association parameter ρ and link function of ρ. $\Phi_2(\cdot,\cdot;\rho)$ denotes the cdf of a standard bivariate normal distribution with correlation coefficient ρ, and $\Phi(\cdot)$ the cdf of a univariate standard normal distribution. Finally, ϵ is set to 10^{-7} and is used to ensure that the restrictions on the space of ρ are maintained.

Copula	$C(u,v;\rho)$	Range of ρ	Link Function
Clayton	$(u^{-\rho}+v^{-\rho}-1)^{-1/\rho}$	$\rho\in(0,\infty)$	$\log(\rho-\epsilon)$
FGM	$uv\{1+\rho(1-u)(1-v)\}$	$\rho\in[-1,1]$	$\tanh^{-1}(\rho)$
Frank	$-\rho^{-1}\log\{1+(e^{-\rho u}-1)\ (e^{-\rho v}-1)/(e^{-\rho}-1)\}$	$\rho\in R\backslash\{0\}$	$-$
Gaussian	$\Phi_2\left(\Phi^{-1}(u),\Phi^{-1}(v);\rho\right)$	$\rho\in[-1,1]$	$\tanh^{-1}(\rho)$
Gumbel	$\exp\left[-\{(-\log u)^\rho+(-\log v)^\rho\}^{1/\rho}\right]$	$\rho\in(1,\infty)$	$\log(\rho-1)$
Joe	$1-\{(1-u)^\rho+(1-v)^\rho-(1-u)^\rho(1-v)^\rho\}^{1/\rho}$	$\rho\in(1,\infty)$	$\log(\rho-1-\epsilon)$

Note that $p_{1,i}$ and $p_{2,i}$ depend on a total of K_1 and K_2 parameters, respectively. In the CGAMLSS approach, all these parameters can be related to an additive predictor, $\eta_i^{\vartheta_k^{(d)}}, k=\{1,\ldots,K_d\}$, $d=1,2$. As in a classical generalized linear regression model a suitable bijective link function, g_k, can be considered which ensures that the restrictions on the parameter spaces are maintained

$$\eta_i^{\vartheta_k^{(d)}}=g_k\left(\vartheta_{ik}^{(d)}\right), k=\{1,\ldots,K_d\}, d=\{1,2\}.$$

The choice of the link function is determined by the restrictions that apply to the parameter space of the corresponding parameter. For example, if one wishes to model the standard deviation as a function of the covariates and regression coefficients, $g_\sigma(\sigma_{1i})=\eta_i^{\sigma_1}$ can be assumed, in which the link function $g_\sigma(\cdot)$ is equal to $\log(\cdot)$ to ensure positive values (see Marra and Radice (2017a) or Klein et al. (2015) for details on the link functions).

Moreover, each copula parameter, say $\vartheta_{i1}^{(c)},\ldots,\vartheta_{iK_c}^{(c)}$, is also related to an additive predictor,

$$\eta_i^{\vartheta_k^{(c)}}, k\in\{1,\ldots,K_c\},$$

by assuming

$$\eta_i^{\vartheta_k^{(c)}}=g_k(\vartheta_{ik}^{(c)}), k\in\{1,\ldots,K_c\}.$$

Note that K_c is the number of copula parameters, in this manuscript, only copulas with one parameter (also denoted by ρ, see Table 4.1) are considered.

In a nutshell, the total number of parameters (K) that can be modelled in the CGAMLSS approach is the sum of the number of parameters of each marginal and the number of parameters of the copula function (i.e. $K=K_1+K_2+K_c$). Let us assume $\boldsymbol{\vartheta}$ the K-dimensional vector formed by all these parameters, $\boldsymbol{\vartheta}=\left(\vartheta_1^{(1)},\ldots,\vartheta_{K_1}^{(1)},\vartheta_1^{(2)},\ldots,\vartheta_{K_2}^{(2)},\vartheta_1^{(c)},\ldots,\vartheta_{K_c}^{(c)}\right)$. All these parameters comprised in $\boldsymbol{\vartheta}$,

are also assumed to be related to regression coefficients and covariates (e.g., binary, continuous and functional) collected in $\boldsymbol{\nu}_i$ via an additive predictor defined as

$$\eta_i^{\vartheta_k} = \beta_0^{\vartheta_k} + \sum_{j=1}^{J} f_j^{\vartheta_k}(\boldsymbol{\nu}_i), \quad k \in \{1, \ldots, K\}, \tag{4.2}$$

where $\beta_0^{\vartheta_k}$ is a general intercept of each predictor, and $f_j^{\vartheta_k}(\boldsymbol{\nu}_i)$ are generic functions of subsets of $\boldsymbol{\nu}_i$, which represent the different covariate effects.

Note that each distribution parameter of marginal distributions and copula function may depend on different covariates and a different number of effects (say J). By dropping the parameter-dependence k and d for notational simplicity, the following generic representation can be used to refer to equation (4.2):

$$\eta_i = \beta_0 + \sum_{j=1}^{J} f_j(\boldsymbol{\nu}_i), \tag{4.3}$$

As in generalized additive regression models each function f_j of equation (4.3) can be modelled by a linear combination of D_j appropriate basis functions

$$f_j(\boldsymbol{\nu}_i) = \sum_{d_j=1}^{D_j} \beta_{j,d_j} B_{j,d_j}(\boldsymbol{\nu}_i). \tag{4.4}$$

Equation (4.4) implies that the vector of evaluations $(f_j(\boldsymbol{\nu}_1), \ldots, f_j(\boldsymbol{\nu}_n))'$ can be written as $\mathbf{Z}_j\boldsymbol{\beta}_j$ with $\boldsymbol{\beta}_j = (\beta_{j,1}, \ldots, \beta_{j,D_j})'$, where $\boldsymbol{\beta}_j$ consists of all the basis coefficients, and the entries $Z_j[i, d_j] = B_{j,d_j}(\boldsymbol{\nu}_i)$ of the design matrix $\boldsymbol{Z}_j$ are the basis functions evaluated at the observed covariate values. Choices of the basis functions depend on the different effect types and we give specific examples at the end of this section.

Finally, Equation (4.3) can be written as

$$\boldsymbol{\eta} = \beta_0\mathbf{1}_n + \mathbf{Z}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{Z}_J\boldsymbol{\beta}_J, \tag{4.5}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'$ represents the predictor vector for all observations, and $\mathbf{1}_n$ is an n-dimensional vector of one.

To regularize the estimation of the potentially high-dimensional vectors of basis coefficients, each vector $\boldsymbol{\beta}_j^{\vartheta_k}$ is associated with a quadratic penalty (in the penalized likelihood framework) or a multivariate Gaussian prior (in the Bayesian framework). More precisely, the quadratic penalties take the form $\lambda\boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta}$ (dropping the parameter index and the function index for simplicity) where the positive semidefinite penalty matrix $\mathbf{K}$ is chosen to enforce the desirable properties of the corresponding functional effect (e.g., smoothness or shrinkage). The smoothing parameter $\lambda \in [0, \infty)$ controls the trade-off between fit and smoothness, and plays a crucial role in the estimation of the shape of the corresponding effect.

In the Bayesian inference model, the penalty term is replaced by a partially improper Gaussian prior

$$p(\boldsymbol{\beta}|\tau^2) \propto \exp(-\frac{1}{2\tau^2}\boldsymbol{\beta}'\boldsymbol{K}\boldsymbol{\beta}), \tag{4.6}$$

where the matrix $\boldsymbol{K}$ is now the prior precision matrix, and τ^2 replaces the smoothing parameter from the penalized likelihood framework. The Bayesian posterior mode and penalized likelihood estimates correspond to each other via $\lambda = \frac{1}{2\tau^2}$.

Many types of effect can be modelled making different assumptions regarding the basis functions and the penalty/prior precision matrix (Wood, 2006; Fahrmeir et al., 2013). The following paragraph discusses only those effects contemplated in the present work.

For linear effects, equation (4.5) can be expressed as $\mathbf{z}'_{ij}\boldsymbol{\beta}_j$ and the design matrix is obtained by stacking all covariate vectors $\mathbf{z}_{ij}$ into $\mathbf{Z}_j$. No penalty is usually assigned to linear effects ($\mathbf{K}_j = \mathbf{0}$) which corresponds to flat priors from a Bayesian point of view. However, penalized parametric effects can sometimes be useful (for example when some factor variables in the model are poorly or not identified by the data).

To render the Bayesian and frequentist modelling comparable in our empirical analysis, the smooth functions of the continuous covariates $f^{\vartheta}(\boldsymbol{\nu})$ were estimated using penalized splines. For Bayesian inference, cubic B-splines with a 20 equidistant inner knot grid were used such that dim($\boldsymbol{\beta}$)=22. The prior for $\boldsymbol{\beta}$ is based on a second or random walk prior with inverse gamma hyperpriors for τ^2 (Lang and Brezger, 2004). For frequentist CGAMLSS, penalized splines with second order penalties were also considered, and the number of basis functions fixed to 20 (as in the Bayesian approach). There are several reasons for the the penalized spline specifications, i) For practical application, the use of penalizations relaxed the election of the number of knots (Rice and Wu, 2001); ii) The number and placement of the knots has only a very minor impact on the fit if the number of knots chosen is not too small (Eilers and Marx, 1996; Lang and Brezger, 2004; Brezger and Lang, 2006); iii) All continuous covariates included in the model have a lot of different values so it makes sense to have a few knots; iv) We have made several tries with different number of knots and 20 equidistant knots yield sufficient flexibility for all the covariates included in the model (data not shown); v) Finally, in Bayesian inference, second order random walk priors leave a linear effect unpenalized which is in analogy to the common penalty for smoothing splines. On the other hand, first order differences often yield more wiggly estimates. In an analogous framework, second order penalizations are considered in frequentist inference.

4.2.2 Likelihood and inference in CGAMLSS

The following lines discuss in more detail both frequentist and Bayesian inferences for continuous-continuous copula regression models.

To simplify the notation, let us suppose that each marginal distribution has two parameters (corresponding to the mean and the scale parameter) and one copula parameter, i.e., $\boldsymbol{\vartheta}_i = (\mu_{1i}, \mu_{2i}, \sigma_{1i}, \sigma_{2i}, \rho_i)$, where μ_{1i} and μ_{2i} are location parameters, σ_{1i} and σ_{2i} are scale parameters of the margins, and ρ_i denotes the association parameter.

Thus, the log-likelihood of a CGAMLSS regression model with continuous margins can be written as:

$$\begin{aligned}\ell(\boldsymbol{\theta}) = &\sum_{i=1}^{n} \log\left\{c(F_1(y_{1i}|\mu_{1i}, \sigma_{1i}), F_2(y_{2i}|\mu_{2i}, \sigma_{2i}); \rho_i)\right\} + \\ &\sum_{i=1}^{n}\sum_{d=1}^{2} \log\left\{p_d(y_{di} \mid \mu_{di}, \sigma_{di}))\right\},\end{aligned} \tag{4.7}$$

for $d = 1, 2$, where $c(\cdot, \cdot, \rho)$ is the density function of the copula function, and $p_d(y_d|\mu_d, \sigma_d)$ the density of the d^{th} marginal. Parameter vector $\boldsymbol{\theta}$ is defined as $(\boldsymbol{\beta}'_{\mu_1}, \boldsymbol{\beta}'_{\mu_2}, \boldsymbol{\beta}'_{\sigma_1}, \boldsymbol{\beta}'_{\sigma_2}, \boldsymbol{\beta}'_{\rho})'$ which refers to the coefficient vectors associated with $\eta_i^{\mu_1}$, $\eta_i^{\mu_2}$, $\eta_i^{\sigma_1}$, $\eta_i^{\sigma_2}$ and η_i^{ρ} respectively.

Bayesian inference

In the CGAMLSS framework, Bayesian inference is carried out using a generic, modular algorithm based on Markov chain Monte Carlo simulations, via the iterative updating of all model parameters of the joint posterior. According to Klein and Kneib (2016b) the log-posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ is given by

$$\log\left((p(\boldsymbol{\theta}|\boldsymbol{y})\right) \propto l(\boldsymbol{\theta}) + \sum_{k=1}^{K}\sum_{j=1}^{J_k} \log(p(\boldsymbol{\beta}_{j,k}|\tau^2_{j,k})p(\tau^2_{j,k})),$$

where $\boldsymbol{\theta}$ is the complete parameter vector (including the smoothing variances) and $\boldsymbol{y}$ denotes the response matrix. This expression is in general intractable for the general types of model considered here. Klein and Kneib (2016b) therefore resorted to a Metropolis-Hastings-algorithm in which proposal densities for the vectors of regression coefficients $\boldsymbol{\beta}_{j,k}$ are obtained by approximating the log-full conditional $\log(p(\boldsymbol{\beta}_{j,k}|.))$ via a second-order Taylor expansion. This yields a multivariate Gaussian proposal density with a mean corresponding to the mode at the current state, and a covariance matrix corresponding to the curvature at this mode. $\boldsymbol{\beta}_{j,k}$ can then be generated via a Metropolis-Hastings update (see, Klein and Kneib, 2016b, for more details).

For the variance parameters $\tau^2_{j,k}$, inverse gamma hyperpriors are assumed ($\tau^2_j \sim IG(a_j, b_j)$, with $a_j = b_j = 0.001$) in order to obtain data-driven smoothness.

Penalized maximum likelihood inference

In the frequentist setting, regression parameter estimations are based on direct optimization of the penalized likelihood with automatic selection of the smoothing parameter. In this type of model, the use of an unpenalized optimization algorithm could produce unduly wiggly estimates - because of the flexible predictor structures employed in a CGAMLSS regression model. Note that penalized approach is a commonly used framework to achieve regularization in models with large numbers of parameters and to achieve smoothness properties of all the function estimates, see Wood (2006) for details.

Marra and Radice (2017a) proposed maximizing the expression

$$l_p(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}'\boldsymbol{S}\boldsymbol{\theta}, \tag{4.8}$$

where $l(\boldsymbol{\theta})$ is the log-likelihood of a CGAMLSS regression model with continuous margins (defined in equation (4.7)); l_p denotes the penalized log-likelihood of the model, and $\mathbf{S} = \text{diag}(\boldsymbol{K}_{\mu_1}, \boldsymbol{K}_{\mu_2}, \boldsymbol{K}_{\sigma_1}, \boldsymbol{K}_{\sigma_2}, \boldsymbol{K}_{\rho})$. The smoothing parameters contained in the $\boldsymbol{K}$ components make up the overall vector $\boldsymbol{\lambda} = (\boldsymbol{\lambda}'_{\mu_1}, \boldsymbol{\lambda}'_{\mu_2}, \boldsymbol{\lambda}'_{\sigma_1}, \boldsymbol{\lambda}'_{\sigma_2}, \boldsymbol{\lambda}'_{\rho})'$ (See Marra and Radice, 2017a, for more details).

Marra and Radice (2017a) proposed estimating $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ using a stable and efficient trust region algorithm with integrated automatic multiple smoothing parameter selection. At convergence, reliable point-wise confidence intervals for linear and non-linear functions of the model coefficients can be obtained using the Bayesian large sample approximation (Marra and Radice, 2017a). See Chapter 6 for a more detailed description of this approach.

A major advantage of the algorithms proposed by Marra and Radice (2017a) and Klein and Kneib (2016b) is that they were created in a modular fashion, and therefore new parametric continuous marginal distributions and copula functions can be easily included. Another is that copula regression parameters are integrated into the estimation of the model coefficients to allow for more flexible dependence modelling. Further, the smoothing parameters are selected automatically.

Software

The reviewed methods can be undertaken using free software. The frequentist approach with CGAMLSS can be performed in R using the `GJRM`-package (Marra and Radice, 2017a). CGAMLSS from a Bayesian perspective can be undertaken using BayesX open-source software (Belitz et al., 2015). The two R-packages `BayesX` (Kneib et al., 2014) and `R2BayesX` (Umlauf et al., 2015; Belitz et al., 2015) can be used to provide graphic interfaces in the Bayesian setting. See Appendix B for more details on how to use it in practice.

In the following, we will compare both approaches via a simulation study.

4.3 Comparison by means of a simulation study

The aim of this simulation study is to compare the two CGAMLSS reviewed methods (Marra and Radice, 2017a; Klein and Kneib, 2016b). The frequentist approach (Marra and Radice, 2017a) will be denoted by `Frequentist`, and the Bayesian by `Bayesian` (Klein and Kneib, 2016b) for the rest of this section. To the best of our knowledge, this is the first time that these methods are compared via simulation studies in the statistical literature.

4.3.1 Scenario 1

Klein and Kneib (2016b) studied the performance of Bayesian CGAMLSS models compared to the penalized likelihood approach VGAM (Yee, 2015). VGAM regression models are available in `VGAM`-package and they allow for the estimation of additive copula models with given marginals. Klein and Kneib (2016b) have shown that `Bayesian` approach outperforms VGAM regression models in terms of accuracy of the estimated copula parameter. Inspired in Klein and Kneib (2016b), we will replicate a similar simulation study, but comparing in this case `Bayesian` and `Frequentist` approaches.

We specify below the parameters of the three models considered in this first scenario

- The sample sizes considered here are $n = 500, 1000,$ and 2000.
- We have considered $R = 1000$ simulation replicates for each sample size.
- For the marginal distributions of the responses, y_1 and y_2, we have considered standard normal margins.
- We have selected three different copulas (Gaussian, Clayton and Gumbel, which are available for both approaches) and their empirical performance will be compared in a simulation study.
- The true effect of the copula parameter was defined as follows

$$f^{\rho}(\nu) = \sin(4\nu) + 2\exp\left(-64\nu^2\right),$$

  where ν is a one continuous covariate i.i.d., defined as $\nu \sim U[-1, 1]$.

- For reasons of simplicity, we have studied the effect of the copula parameter conditioned by covariates (as in Klein and Kneib (2016b)). Consequently, the predictor equations for y_1 and y_2 and the copula parameter are given by

$$\begin{cases} \eta_i^{\mu_1} = \eta_i^{\sigma_1^2} = \eta_i^{\mu_2} = \eta_i^{\sigma_2^2} = 1 \\ \eta_i^{\rho} = \beta_0^{\rho} + f^{\rho}(\nu). \end{cases}$$

- The copula parameters are specified as follows,
  - For **Gaussian copula** and `Bayesian` software, ρ is set as $\rho = \frac{f(\nu)}{\sqrt{1+(f(\nu))^2}}$ by default. For `Frequentist`, Marra and Radice (2017a) shows that the transformation of ρ is given by $\tanh^{-1}(\rho)$, i.e, $f(\nu) = tanh^{-1}(\rho) \Rightarrow \rho = \tanh(f(\nu)) = \frac{e^{\nu}-e^{-\nu}}{e^{\nu}+e^{-\nu}} = \frac{e^{-\nu}\left(e^{2\nu}-1\right)}{e^{-\nu}(e^{2\nu}+1)} = \frac{e^{2\nu}-1}{e^{2\nu}+1}$. The copula parameter in `Frequentist` is then specified as $\rho = \tanh(f(\nu)) = \frac{e^{2\nu}-1}{e^{2\nu}+1}$, i.e, the same than in VGAM (Yee and Wild, 1996). To make the results comparable, ρ is set to $\tanh^{-1}(\rho)$ for both approaches, `Frequentist` and `Bayesian`.
  - For **Gumbel copula**, the link used here is a shift log-link for both approaches (by default) so $\rho = \exp(f(\nu)) + 1$.
  - For **Clayton copula**: For this copula, $\rho = \exp(f(\nu))$ is considered (by default) for both approaches.

- To render `Frequentist` and `Bayesian` comparable in our simulation study, the smooth functions of the continuous covariates were estimated using P-splines. For Bayesian inference, cubic B-splines with a 18 equidistant inner knot grid were used such that dim($\boldsymbol{\beta}$)=22. The prior for $\boldsymbol{\beta}$ is based on a second order random walk prior with inverse gamma hyperpriors for τ^2: $a = b = 0.001$ (Lang and Brezger, 2004) - the option by default for `Bayesian`. For `Frequentist`, penalized splines with second order penalties were also considered, and the number of basis functions fixed to 20 (as in the Bayesian approach).

- To evaluate the simulation study, logarithmic mean squared errors (MSEs), pointwise standard errors and coverage rates of 95% pointwise credible intervals were considered here.

Results are summarized in Figure 4.2-4.5 and Table 4.2. Figure 4.2 shows the logarithmic mean squared error (MSE), of the estimated copula parameters obtained for both approaches `Bayesian` and `Frequentist`. Figures 4.3-4.5 represents non-linear estimates for each approach. Time computing was also compared in Table 4.2. The main conclusions of the simulated study are described below.

- As expected, MSE is smaller at higher samples sizes. Note that the differences observed in MSE with both approaches are less than 0.011 for all sample sizes considered (see Figure 4.2). Indeed, both approaches captured the true effect of the copula parameters simulated (see Figures from 4.3 to 4.5).

- It can be shown that Gumbel and Clayton copula parameters are more difficult to estimate for $n = 500$ than $n = 1000$, and 2000. See Figures from 4.3 to 4.5.
- In general, Clayton parameters are more difficult to estimate than Gaussian or Gumbel parameters for both approaches. See Figures from 4.3 to 4.5.
- Note that in this simulation study we are using P-splines for both approaches, `Frequentist` and `Bayesian` to make the results comparable. However, for `Frequentist` the smooth option by default is thin plate regression splines. It could justify the picks observed in some replicates, specially for Clayton copula parameters.
- Finally, we have compared computing times of the two approaches. See Table 4.2. We have found that `Frequentist` is faster than `Bayesian` approach. However, we should note that mean time computing of each iteration is less than one and a half minute for all the models considered here.
- As expected, time computing increases as the sample size also increases. Furthermore, time computing is higher for Gumbel and Clayton copulas than for the Gaussian. See Table 4.2.

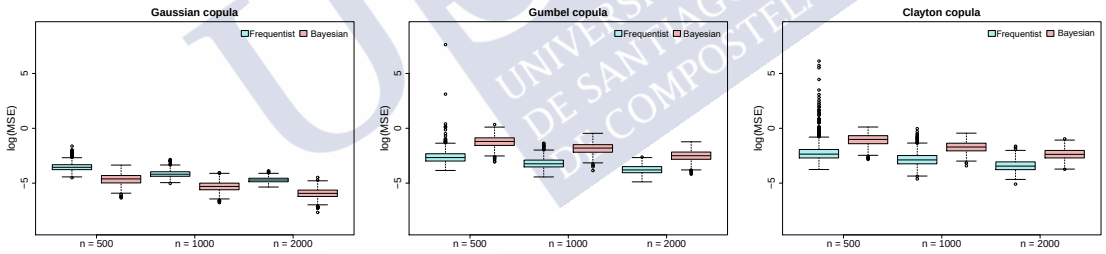


Figure 4.2: Logarithmic mean squared errors obtained by applying `Frequentist` and `Bayesian` approaches to data simulated from scenario 1.

4.3.2 Scenario 2

We will also study the performance of `Bayesian` and `Frequentist` approaches in a second scenario. In this study, we will replicate a similar simulation study that the presented in Marra and Radice (2017a), but comparing in this case the `Bayesian` and `Frequentist` approaches. We specify below the parameters of the three models considered here:

- The sample sizes are $n = 500, 1000,$ and 2000 with $R = 1000$ simulation replicates for each sample size.

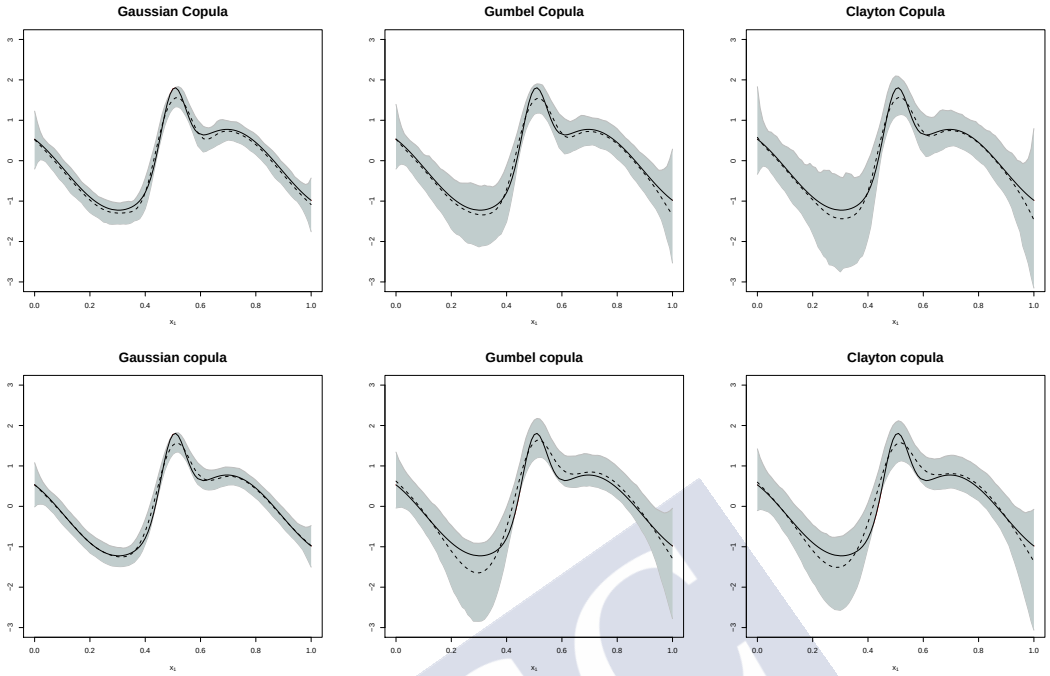


Figure 4.3: Smooth function estimates obtained with `Frequentist` (first row of the panel) and `Bayesian` (second row) approaches to data simulated from **n=500** of **scenario 1**. The true function was represented by black solid lines, mean estimates by dashed lines and pointwise ranges resulting 5% and 95% quantiles from the 1000 replications by shaded areas.

- For the marginal distributions of the responses, y_1 and y_2, we have considered standard normal margins.

- We selected three different copulas (Gaussian, Clayton and Gumbel, which are available for both approaches) and their empirical performance will be compared in a simulation study.

- The true effect of the copula parameter were defined as follows,

$$f^{\rho}(\nu) = \sin(2\pi\nu),$$

  where ν is a one continuous covariate i.i.d., defined as $\nu \sim U[-1, 1]$.

- The link copula parameters were specified as in scenario 1.

- The smooth functions of the continuous covariates were estimated as in scenario 1.

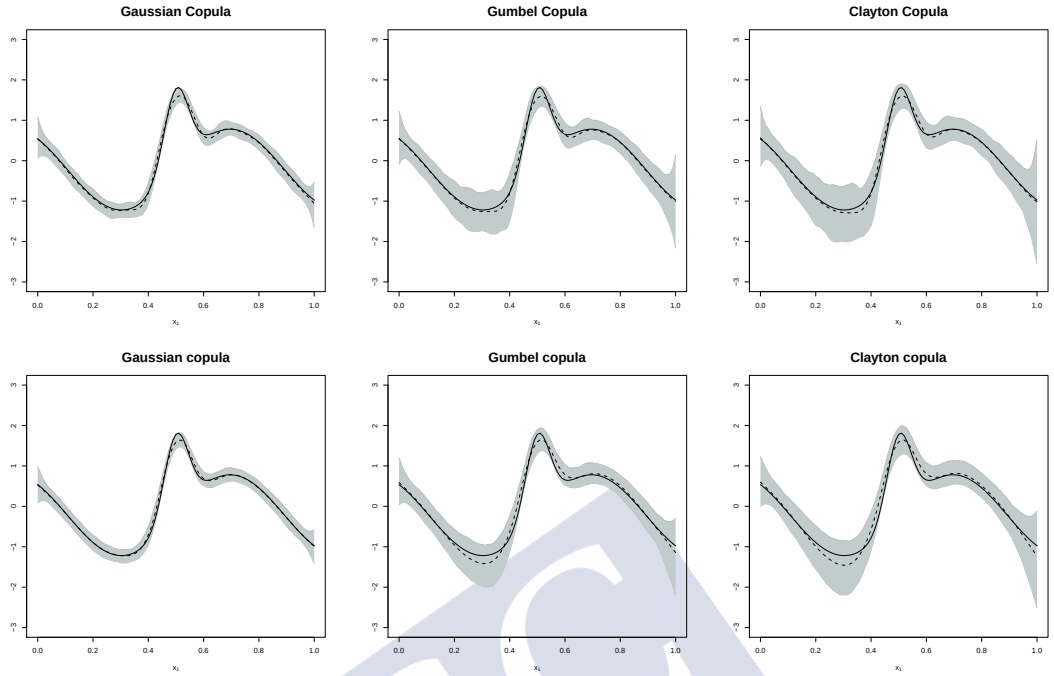


Figure 4.4: Smooth function estimates obtained with `Frequentist` (first row of the panel) and `Bayesian` (second row) approaches to data simulated from **n=1000** of **scenario 1**. The true function was represented by black solid lines, mean estimates by dashed lines and pointwise ranges resulting 5% and 95% quantiles from the 1000 replications by shaded areas.

- The predictor equations for y_1 and y_2 and the copula parameter are given by

$$\begin{cases} \eta_i^{\mu_1} = \eta_i^{\sigma_1^2} = \eta_i^{\mu_2} = \eta_i^{\sigma_2^2} = 1 \\ \eta_i^{\rho} = \beta_0^{\rho} + f^{\rho}(\nu). \end{cases}$$

- Again, logarithmic MSE of the copula parameter estimates, were considered to study the performance of both approaches.

Results are summarized in Figures 4.6-4.9 and Table 4.3. As can be seen in the figures, estimations are excellent with both approaches. In the following, we summarized the main conclusions of the results obtained

- As expected, all copula parameters estimated are very close to the true effect and variability decreases at higher sample size.
- In terms of MSE, there is no differences between `Frequentist` and `Bayesian` approaches.
- Gaussian copula parameters are easier to estimate than Gumbel and Clayton copula parameters, specially for $n = 500$ for both approaches.

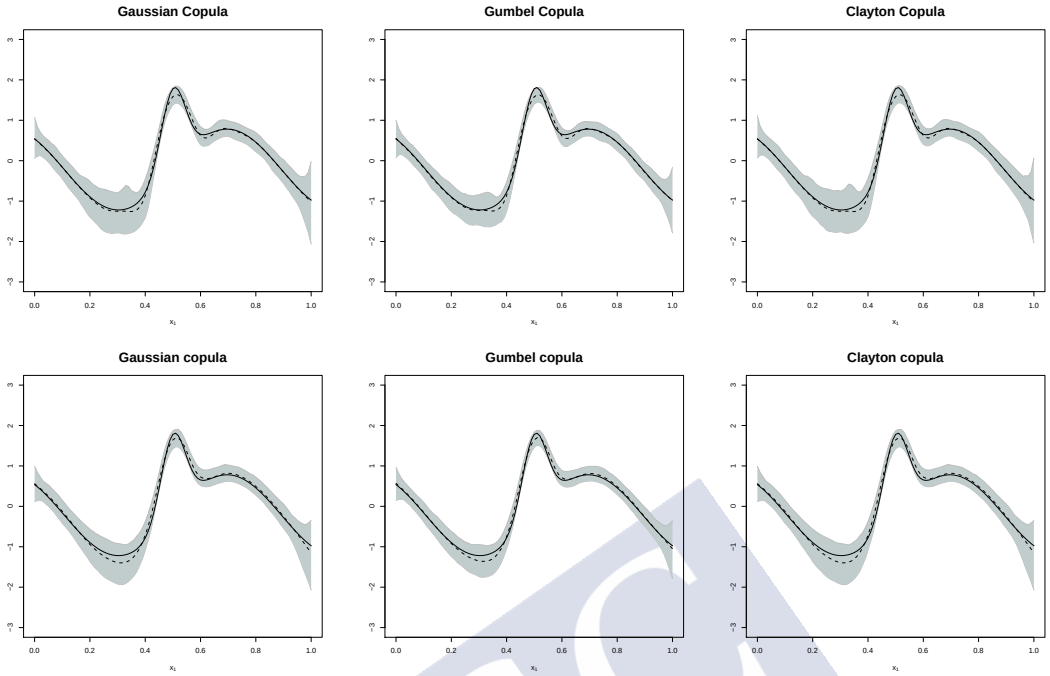


Figure 4.5: Smooth function estimates obtained with `Frequentist` (first row of the panel) and `Bayesian` (second row) approaches to data simulated from **n=2000** of **scenario 1**. The true function was represented by black solid lines, mean estimates by dashed lines and pointwise ranges resulting 5% and 95% quantiles from the 1000 replications by shaded areas.

- Concerning time computing, results are also similar that the ones obtained in scenario 1 (see Table 4.3). We found that `Frequentist` is faster than `Bayesian`. It should be noted that computing is fast for both approaches. Specifically, mean time computing for each iteration and scenario is less than one and a half minute for all the models considered here.

- As expected, time computing increases as the sample size increases. Furthermore, time computing is higher for Gumbel and Clayton copulas comparing with the Gaussian. See Table 4.3.

4.4 Joint modelling of glycation data

4.4.1 The A-Estrada Glycation and Inflammation Study (AEGIS)

AEGIS is a cross-sectional, population-based study being performed in the municipality of A Estrada (NW, Spain). Its aim is to investigate the association between glycation, inflammation, lifestyles and their association with common

Table 4.2: Mean time computing values (in seconds) of each replicate for scenario 1 when using a 3.6-GHz Intel(R) Core(TM) i7-7700 running Linux. In this table, `Freq.` denotes the `Frequentist` approach (Marra and Radice, 2017a) and `Bayes.` the `Bayesian` approach (Klein and Kneib, 2016a).

	n=500		n=1000		n=2000	
Copula	`Bayes.`	`Freq.`	`Bayes.`	`Freq.`	`Bayes.`	`Freq.`
Gaussian	6.94	0.68	12.58	0.90	24.48	1.23
Gumbel	23.57	1.24	45.42	1.60	90.00	2.79
Clayton	10.40	1.20	20.17	1.30	39.75	1.96

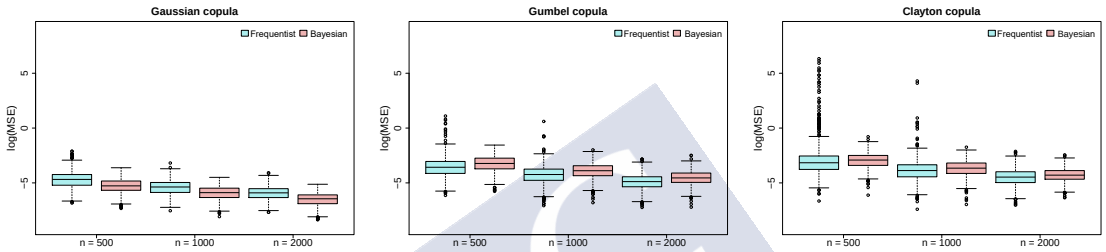


Figure 4.6: Logarithmic mean squared errors obtained by applying `Frequentist` (in blue) and `Bayesian` approaches (in pink) to data simulated from scenario 2.

diseases, and to study discordances between markers for glycaemia (Gude et al., 2017). An outline of the AEGIS study is available at www.clinicaltrials.gov, code NCT01796184. The study was reviewed and approved by the Clinical Research Ethics Committee from Galicia, Spain (CEIC2012-025). Written informed consent was obtained from each participant in the study, which conformed to the current Helsinki Declaration.

An age-stratified random sample of the population aged 18 years and older was drawn from Spain's National Health System (NHS) Registry, which covers more than 95% of the population and contains the name, birth date and address of every person entitled to NHS primary care. The sample was stratified into

Table 4.3: Mean time computing values (in seconds) of each replicate for scenario 2 when using a 3.6-GHz Intel(R) Core(TM) i7-7700 running Linux. In this table, `Freq.` denotes the `Frequentist` approach (Marra and Radice, 2017a) and `Bayes.` the `Bayesian` approach (Klein and Kneib, 2016a).

	n=500		n=1000		n=2000	
Copula	`Bayes.`	`Freq.`	`Bayes.`	`Freq.`	`Bayes.`	`Freq.`
Gaussian	6.44	0.82	12.44	1.01	24.43	1.55
Gumbel	22.94	1.27	45.2	1.99	89.68	2.95
Clayton	10.21	1.36	19.95	1.72	39.36	2.22

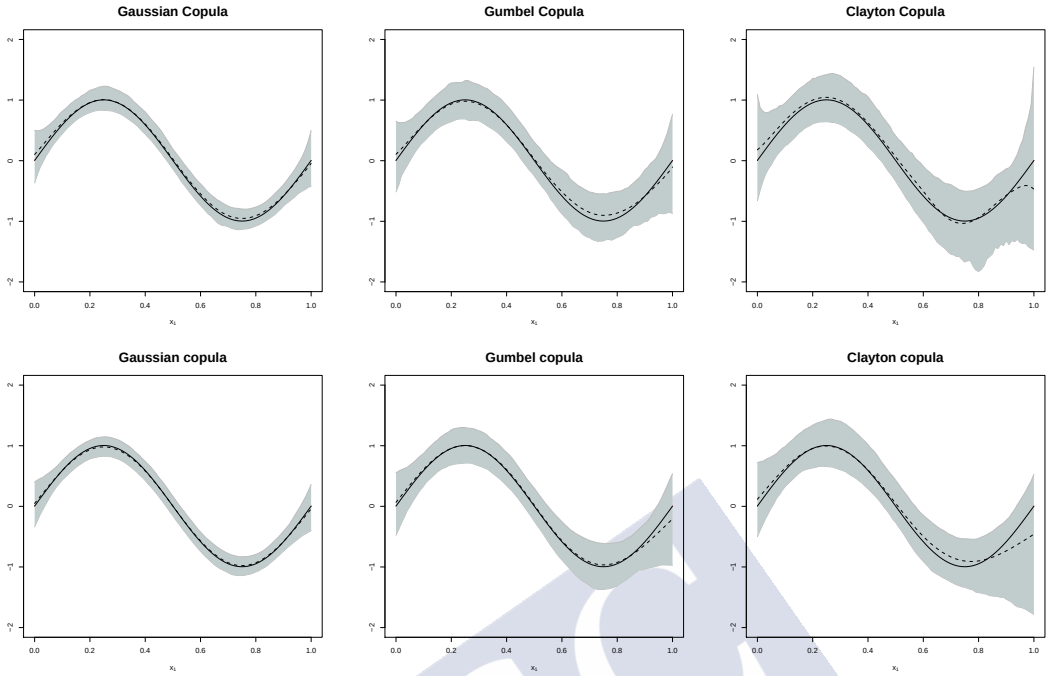


Figure 4.7: Smooth function estimates obtained with `Frequentist` (first row of the panel) and `Bayesian` (second row) approaches to data simulated from **n=500** of **scenario 2**. The true function was represented by black solid lines, mean estimates by dashed lines and pointwise ranges resulting 5% and 95% quantiles from the 1000 replications by shaded areas.

the following age groups: 18 to 29 years; 30 to 39 years; 40 to 49 years; 50 to 59 years; 60 to 69 years; 70 to 79 years; and 80 years and older. A computer program generated a random sample of equal numbers ($n = 500$) of subjects in each age group. Of this initial sample of 3500 individuals, the following were excluded: 428 due to completion of the recruitment period; 84 due to death; 211 due to non-response; and 134 due to change of address. Furthermore, any subjects who could not provide written informed consent was deemed ineligible to participate in the study; this group included patients with dementia, mental retardation, cerebrovascular disease, terminal cancer, or inability to communicate. Of the remaining eligible 2624 persons, a total of 1516 subjects agreed to participate in the study (overall participation rate, 68%). Participation was lower, not only among men (65%) versus women (71%), but also in the youngest and oldest age groups. There were no significant differences in terms of age or residence (rural/urban) between subjects who did and did not agree to participate in the study. From November 2012 through March 2015, all subjects were successively convened for one day at the Primary Care Center for evaluation, which included an interviewer-administered structured questionnaire and fasting venous blood sampling (Gude et al., 2017).

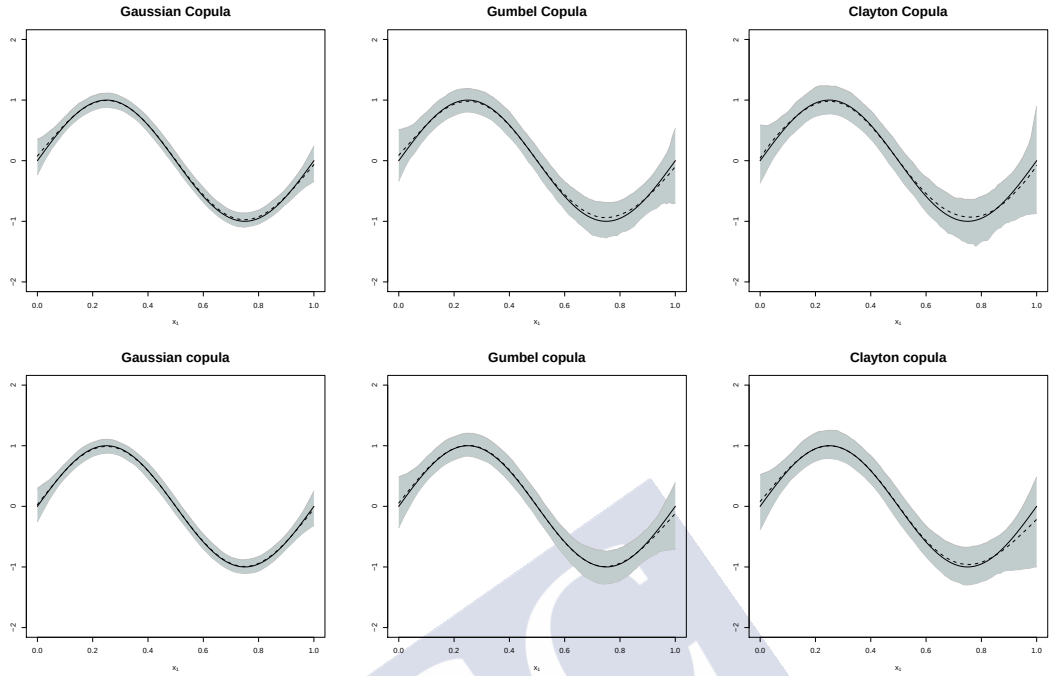


Figure 4.8: Smooth function estimates obtained with `Frequentist` (first row of the panel) and `Bayesian` (second row) approaches to data simulated from **n=1000** of **scenario 2**. The true function was represented by black solid lines, mean estimates by dashed lines and pointwise ranges resulting 5% and 95% quantiles from the 1000 replications by shaded areas.

Data description

The participants mean age was 52 years (range 18 to 91), 55% were females, and 187 (12%) had been previously diagnosed with diabetes. Among those with diabetes, 66.8% took oral anti-diabetics, 3.7% took insulin alone, and 13.3% took insulin plus oral drugs. The remaining 16.2% took none of these medications. Participants with elevated HbA1c and fructosamine levels were more likely to be older, to have fewer years of education, and were more likely to be current smokers than to have formerly used or never used tobacco. They were also less likely to undertake health enhancing physical activity and to be alcohol drinkers. HbA1c and fructosamine concentrations were highly correlated (Pearson correlation coefficient, r = 0.72), and the concentrations of both proteins correlated with fasting plasma glucose levels (r = 0.72 and r = 0.56 respectively).

In determining the factors that influence the HbA1c and fructosamine concentrations, their variability, and the relationship between them, the following variables were deemed covariates: fasting plasma glucose (glucose in formulae, mg/dL), age (in years), gender, body mass index (BMI, Kg/m^2), plasma albumin (Alb in formulae, f/L), and the mean corpuscular (red blood cell) volume

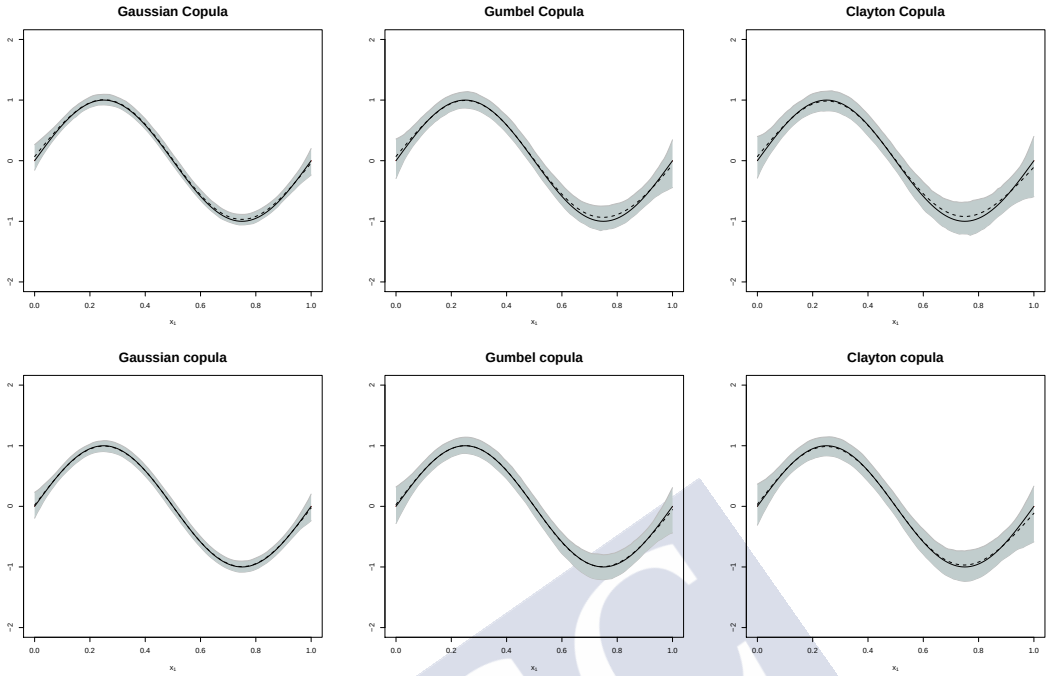


Figure 4.9: Smooth function estimates obtained with `Frequentist` (first row of the panel) and `Bayesian` (second row) approaches to data simulated from **n=2000** of **scenario 2**. The true function was represented by black solid lines, mean estimates by dashed lines and pointwise ranges resulting 5% and 95% quantiles from the 1000 replications by shaded areas.

(MCV, fL). The HbA1c and fructosamine concentrations were considered to be the bivariate response. All laboratory analyses were performed on the day of sample collection in the Clinical Biochemistry Laboratory of the Hospital Clínico Universitario de Santiago de Compostela, Spain. A more detailed description of participant's clinical characteristics and laboratory measurements is given in Table 4.4.

4.4.2 Model building

This section describes the construction of a bivariate model for studying the relationship between HbA1c (Y_1) and fructosamine (Y_2). This involved making the choice of appropriate marginal distributions and a suitable copula function. The Akaike and Bayesian Information Criteria (AIC/BIC) can be used to deal with model selection in the frequentist setting (Marra and Radice, 2017a), while the Deviance Information Criterion (DIC, Spiegelhalter et al., 2002) and the Widely Applicable Information Criterion (WAIC, Watanabe, 2013) can be used to choose the best response distributions and a suitable copula function in the Bayesian framework. The DIC is a commonly used criterion for model choice in Bayesian

Table 4.4: Participant's clinical characteristics according to three different glycemic status: Normo-glycaemic (FPG < 100 mg/dL or HbA1c $< 5.7\%$); Prediabetes (100 mg/dL $\leq$ FPG ≤ 125 mg/dL or $5.7\% \leq$ HbA1c $< 6.5\%$); Diabetes (HbA1c $\geq 6.5\%$ or FPG > 125 mg/dL). Continuous variables are summarize in terms of means $\pm$ standard deviation. Categorical variables are presented as absolute frequency (%). Here, FPG denotes Fasting Plasma Glucose, MCV (Mean Corpuscular Volume) and BMI (Body Mass Index). *Physical activity* was evaluated using *The International Physical Activity Questionnaire* (Craig et al., 2003). The questionnaire records the time spent on different type of activities weighted according to some resting metabolic rates. Subjects were classified into three levels: *inactive, minimal active* and *"HEPA active"* (health-enhancing physical activity, the highest active category). Overweight ranging from a BMI of 25 kg/m^2 to 30 kg/m^2, *Obese*: BMI ≥ 30 kg/m^2; *Normal weight*: BMI < 25 kg/m^2. Alcohol consumption was measured using the standard drinking unit system (see Gual et al., 1999). Individuals were classified into four categories according to their alcohol consumption: *abstainers* (individuals with a regular alcohol consumption of 0 g per week); *light drinkers* (alcohol consumption between 1 g to 139 g per week); *moderate drinkers* (alcohol consumption between 140 g to 279 g per week) and *heavy drinkers* (alcohol consumption ≥ 280 g per week). Tobacco consumption was assessed trough the number of cigarettes usually consumed per day, patients who smoke at least one cigarette by day or quit smoking during the previous year has been considered *smokers*.

Variable	Normo-glycaemic (n=1134)	Prediabetes (n=267)	Diabetes (n=115)	Overall sample (n=1516)
Age, years	49 ± 17	63 ± 13	64 ± 13	52 ± 17
Gender				
Female	658 (58%)	131 (49%)	49 (43%)	838 (55%)
Male	476 (42%)	136 (51%)	66 (57%)	678 (45%)
BMI				
Normal weight	386 (34%)	26 (10%)	26 (10%)	423 (28%)
Overweight	447 (39%)	90 (34%)	37 (32%)	574 (38%)
Obese	301 (27%)	151 (56%)	67 (58%)	519 (34%)
Physical activity				
Inactive	414 (37%)	125 (47%)	57 (50%)	596 (39%)
Minimally active	431 (38%)	81 (30%)	40 (35%)	552 (36%)
"HEPA active"	289 (25%)	61 (23%)	18(15%)	368 (24%)
Alcohol consumption				
Abstainers	426 (38%)	80 (30%)	40(35%)	546(36%)
Light drinkers	479 (42%)	86 (32%)	33 (29%)	598 (39%)
Moderate drinkers	149 (13%)	67 (25%)	25 (22%)	241 (16%)
Heavy drinkers	80 (7%)	34 (13%)	17 (14%)	131 (9%)
Smoking				
Non-smokers	598 (62%)	166 (53%)	166(53%)	825 (54%)
Ex-smokers	276 (30%)	79 (35%)	79 (35%)	395 (26%)
Smokers	260 (8%)	61 (23%)	22 (12%)	296 (20%)
FPG, mg/dL	85 ± 8	108 ± 7	157 ± 31	94 ± 23
HbA1c, %	5.4 ± 0.4	5.9 ± 0.6	7.3 ± 1.0	5.6 ± 0.7
Fructosamine, μ mol/L	248 ± 44	272 ± 64	375 ± 95	262 ± 63
MCV, f/L	89.6 ± 4.7	90.4 ± 4.8	90.2 ± 5.5	89.9 ± 4.8
Albumin, f/L	4.4 ± 0.2	4.4 ± 0.2	4.4 ± 0.2	4.4 ± 0.2

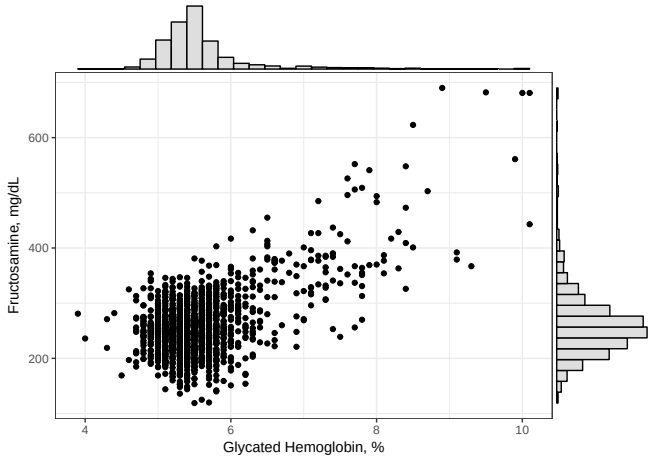


Figure 4.10: Relationship between fructosamine and glycated haemoglobin (HbA1c). The results of some descriptive statistics for fructosamine are 144 (Min.), 225 (1st Qu.), 252 (Median), 262.5 (Mean), 280 (3rd Qu.) and 690 (Max.). Those for HbA1c are 3.9 (Min.), 5.2 (1st Qu.), 5.4 (Median), 5.5 (Mean), 5.6 (3rd Qu.) and 10.1 (Max.).

regression models. It became popular partly because of its easy implementation from the Markov chain Monte Carlo (MCMC) output. The performance of the DIC was evaluated positively by Klein and Kneib (2016b), who compared several misspecified models with the true model using this criterion.

In addition, Marra and Radice (2017a) showed via simulation studies that in the frequestist framework AIC and BIC are able to identify the correct copula function and hence the correct type of dependence structure.

It could be noted hat if the user chooses the wrong marginal distribution this can also affect the choice of the copula but this will depend on the severity of marginal misspecification. To avoid this situation, we propose to start with the selection of two adequate marginal distributions that fit each response satisfactorily as two different and independent GAMLSS regression models. In this univariate framework, model selection was extensively studied in the statistical literature. See, for example, Klein et al. (2015) for a detailed guide for dealing with model choice.

Marginal distributions

AIC/BIC (in the frequentist approach) and WAIC/DIC (in the Bayesian approach) showed log-normal distributions to provide the best fit for both margin distributions within a set of candidate distributions (data not shown). The log-normal distribution is characterized by a location parameter μ plus a scale parameter σ. For a log-normally distributed random variable y with a density function

$f(y|\mu,\sigma) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left[-\frac{\{\log(y)-\mu\}^2}{2\sigma^2}\right]$, the expectation and the variance can be expressed as $\mathbb{E}(y) = e^{\frac{\sigma^2}{2}} e^{\mu}$ and $\mathrm{Var}(y) = e^{\sigma^2}\left(e^{\sigma^2}-1\right)e^{2\mu}$.

Additive predictors

In CGAMLSS, the analyst has to define regression predictors for each parameter of the response distribution. Good knowledge of the biological process, plus the information criteria, can be used to guide the selection of the covariates for each predictor component. For example, in this study MCV was selected as a covariate in the first marginal (HbA1c), and albumin in the second (fructosamine). The MCV is the mean volume of red blood cells, and is useful in classifying the type of anaemia based on red cell morphology. A higher than normal MCV indicates that the red blood cells are too big, and could reflect folic acid deficiency, vitamin B12 deficiency, or alcohol consumption. Conversely, a low MCV reveals the volume of red blood cells to be below normal - a very common condition that usually reflects iron deficiency, especially in women with heavy menstrual bleeding. In both of these anaemia conditions, the blood circulation time of the red cells is reduced, and since haemoglobin is found inside red cells, the lifespan of the red cells could be reduced too. Serum albumin is the main substrate to which glucose binds to, forming serum fructosamine.

Age and BMI were included as covariates since they can be expected to modify the glycation processes (Pani et al., 2008; Huh et al., 2014). For the scale parameter, age and glucose were considered since they are involved in the variability of the glycation rate (see Figure 4.10). Glucose, age and MCV were contemplated as covariates for the same reasons. In addition, values for AIC/BIC and DIC/WAIC were taken into account when selecting the final model. Finally, the additive predictors for the parameters of the joint distribution were specified as

$$\begin{cases} \eta_i^{\mu_1} = \beta_0^{\mu_1} + f_1^{\mu_1}(Glucose_i) + f_2^{\mu_1}(Age_i) + Gender_i\beta_{1i}^{\mu_1} + f_3^{\mu_1}(BMI_i) + f_4^{\mu_1}(MCV_i) \\ \eta_i^{\sigma_1^2} = \beta_0^{\sigma_1^2} + f_1^{\sigma_1^2}(Glucose_i) + f_2^{\sigma_1^2}(Age_i) + Gender_i\beta_{1i}^{\sigma_1^2} \\ \eta_i^{\mu_2} = \beta_0^{\mu_2} + f_1^{\mu_2}(Glucose_i) + f_2^{\mu_2}(Age_i) + Gender_i\beta_{1i}^{\mu_2} + f_3^{\mu_2}(BMI_i) + f_4^{\mu_2}(Alb_i) \\ \eta_i^{\sigma_2^2} = \beta_0^{\sigma_2^2} + f_1^{\sigma_2^2}(Glucose_i) + f_2^{\sigma_2^2}(Age_i) + Gender_i\beta_{1i}^{\sigma_2^2} \\ \eta_i^{\rho} = \beta_0^{\rho} + f_1^{\rho}(Glucose_i) + f_2^{\rho}(Age_i) + Gender_i\beta_{1i}^{\rho} + f_3^{\rho}(MCV_i). \end{cases} \tag{4.9}$$

Hereafter the parameter index is eliminated for the sake of simplicity. The predictors (η_i) are formed through the additive composition of an intercept β_0 representing the overall level of the predictor, linear effects for gender, and functions f reflecting the non-linear effects of the continuous covariates (glucose, age, BMI, albumin and MCV). The first and third equations of equation (4.9) refer to the location parameters μ_1 and μ_2 of HbA1c and fructosamine respectively, while the second and fourth equations refer to the scale parameters σ_1^2 and σ_2^2. The eighth equation refers to the association parameter of the copula ρ.

Selection of copula function

As for the choice of the copula, different copula functions available in both the frequentist and the Bayesian model formulations were contemplated. In this work, we focus on the Gaussian, Gumbel and Clayton copulas to make our analysis concise and justified. These represent the classes of copulas with no (Gaussian), upper (Gumbel) and lower (Clayton) tail dependence as yielded the best results in terms of our model selection criteria and with respect to numerical stability and convergence. The best fit was provided by the Gumbel copula in both frameworks (frequentist and Bayesian). See Table 6.2.

Table 4.5: Comparison of model choice criteria under different copula assumptions.

	Frequentist		Bayesian	
Copula	AIC	BIC	DIC	WAIC
Gumbel	16008.93	16328.37	5905.20	5960.84
Gaussian	16073.56	16486.07	5939.74	5993.23
Clayton	17466.43	18185.62	6047.67	6130.85

Note in that respect that while it is often possible to identify whether the conditional responses between these three types of dependence structures (no, upper, and lower tail dependence) it is sometimes hard to identify the best copula within one of these classes. We should also note that, the software BayesX does not allow for estimation of all the copulas that are currently available in GJRM. To a more detailed study of the dependence structure, we estimated the models also with further copulas implemented in the GJRM package and report the results in Table 4.6. We have observed numerical issues with the Joe copula. But as can be seen in the table, Frank and FGM copula yield higher AIC/BIC values than the Gumbel copula such that the copula choice does not change.

Residuals can be used to check the performance of the selected model. If the estimated model is close to the true model, the normalized quantile residuals approximately follow a Gaussian distribution. Note that the residual checks are only for the margins and not the whole model. Figure 4.11 shows the histograms, density estimates and quantile-quantile plots for the margin models for

Table 4.6: Comparison of model choice criteria under different copula assumptions using GJRM.

	Copula model selection		
Copula	AIC	BIC	Convergence Warnings
Joe	113456.20	113663.10	Warnings obtained
Frank	16181.07	16530.73	No Warnings obtained
FGM	16194.27	16588.31	No Warnings obtained

HbA1c (top) and fructosamine (bottom). Reference bands are included for judging the departures of the quantile-quantile plots from the ideal (red line). The log-normal distribution appears appropriate for residuals in the range -2.5 to 2.4 for both margin models, but deviates from the diagonal for extreme values. This was the best-fitting distribution.

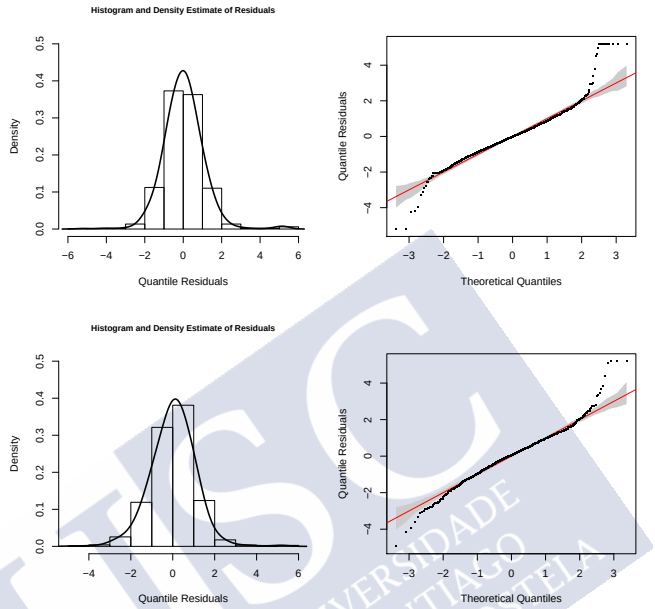


Figure 4.11: Histograms and quantile-quantile plots of normalized quantile residuals for glycated haemoglobin (top) and fructosamine (bottom) for the selected model. The closer the residuals to the bisecting line, the better the fit to the data. Note that residuals are only indicating the goodness of fit in the marginals.

The non-linear effects of continuous variables are typically represented as centred in additive regression models. It should be noted that, in this study, new functions were developed to plot these effects on the real scale for each parameter of the bivariate responses. Figures 4.12, 4.13 and 4.14 show the results obtained. Variability is expressed as the standard deviation, and the association parameter with Kendall's τ. Given that the magnitudes of the copulas' dependence parameters (ρ) are not comparable between copulas, it is normal to use association measurements such as Kendall's τ to facilitate interpretation. Kendall's τ is a well known statistical coefficient that allows one to study the strength of the relationship between two variables (Joe, 1997; Nelsen, 2006). Further, for each copula function a relation exists between ρ and Kendall's $\tau \in [-1, 1]$. CGAMLSS regression models include link functions to ensure that the restrictions on the parameters' spaces are maintained. Specifically, for the copula parameter it can be shown that the following logarithmic link function can be used in the Gumbel case

Table 4.7: Summary of estimated linear effects for model (4.9) obtained from BayesX software. The results were analogous in the frequentist framework (data not shown).

Parameter	mean	2.5% quantile	median	97.5% quantile
$\beta_0^{\mu_1}$ (intercept)	1.01	0.91	1.01	1.11
$\beta_0^{\sigma_1^2}$ (intercept)	-0.71	-0.89	-0.71	-0.52
$\beta_1^{\mu_1}$ (gender)	-0.04	-0.09	-0.04	0.02
$\beta_1^{\sigma_1^2}$ (gender)	0.32	0.17	0.33	0.47
$\beta_0^{\mu_2}$ (intercept)	0.84	0.71	0.84	1.00
$\beta_0^{\sigma_2^2}$ (intercept)	-0.27	-0.45	-0.28	-0.11
$\beta_1^{\mu_2}$ (gender)	-0.26	-0.34	-0.26	-0.18
$\beta_1^{\sigma_2^2}$ (gender)	0.04	-0.11	0.04	0.19
β_0^{ρ} (intercept)	-0.18	-2.37	-1.82	-1.33
β_1^{ρ} (gender)	0.14	-0.41	0.14	0.67

$$\log(\rho - 1) = \eta^{\rho},\ \rho \in (1, \infty) \Longleftrightarrow \rho = \exp(\eta^{\rho}) + 1, \tag{4.10}$$

i.e., in our case

$$\log(\rho - 1) = \eta^{\rho} = \beta_0^{\rho} + f^{\rho}(Glucose) + f^{\rho}(Age) + Gender\beta_1^{\rho} + f^{\rho}(MCV), \rho \in (1, \infty).$$

On the other hand, for each copula function a relation exists between ρ and Kendall's τ. In the case of considering the Gumbel copula, it can be shown that

$$\tau = 1 - \frac{1}{\rho}.$$

Further it is well known that Kendall's τ takes values from -1 to 1. The use of the link function defined in (4.10) ensures that Kendall's τ will be estimated correctly since ρ is always higher than 1. Moreover, for the specific Gumbel copula Kendall's τ takes values from 0 to 1. Therefore, Kendall's τ is correctly defined and η^{τ} takes the form

$$\eta^{\tau} = 1 - \frac{1}{\exp(\eta^{\rho}) + 1}.$$

4.4.3 Results

Before describing the results obtained, it should be understood that the Bayesian and frequentist approaches returned very similar results. Thus, a single set of results are presented.

Marginal expectations

Gender had no influence on the mean concentration of HbA1c. However, it had some influence on the fructosamine concentration, although this would seem to be clinically irrelevant -the differences between men and women were minimal (smaller than 0.1 mg/dL). See, Table 4.7.

Fasting plasma glucose was the main covariate influencing the concentrations of both HbA1c and fructosamine. The functional form of the effect of glucose levels on these proteins was similar (Figures 4.12). The relationship between HbA1c and glycaemia has been extensively explored in studies by other authors, the results of which support an association between HbA1c and the glucose concentration during the preceding 5-12 weeks (Koenig et al., 1976; Nathan et al., 2007).

The mean HbA1c concentration increased almost linearly with age, but the fructosamine concentration only did so in elderly people ($>$ 50 years). These findings are consistent with the view that glycation is accelerated by ageing (Davidson, 1979). The age-related increase in HbA1c is similar in magnitude to that reported in the Framingham Offspring Study (FOS), which examined data from 2473 non-diabetic participants, as well as that reported in the National Health and Nutrition Examination Survey NHANES, 2001-2004 which involved 3270 non-diabetic participants. Since the HbA1c concentration increased with age after adjusting for glucose levels, factors unrelated to glucose metabolism must be involved. One such factor may be the ageing-related change in the rate of glycation. Pani et al. (2008) also reported clear differences between HbA1c and fructosamine concentrations in subjects of different ages.

In individuals suffering from overweight or obesity (BMI $>$ 25 Kg/m^2 and BMI $>$ 30 Kg/m^2 respectively), higher BMI values were associated with a higher mean concentration of HbA1c and a lower mean concentration of fructosamine. Several studies have suggested a negative correlation between BMI and serum glycated proteins in people with and without diabetes (Miyashita et al., 2007; Koga et al., 2006; Huh et al., 2014). Some authors suggest that the inverse association between extra-intravascular glycated proteins and BMI is the result of different mechanisms coming into play depending on the glucose tolerance status. In people without diabetes it would appear to be due to a direct association between BMI and glycated proteins, while in people with diabetes, glycated proteins are influenced by plasma glucose values as well (Huh et al., 2014).

Both higher and lower MCVs appear associated with lower levels of glycated haemoglobin. It is well known that the formation of HbA1c increases in erythrocytes over their lifetime: younger cells contain smaller amounts and older cells larger amounts. Since the circulation time of the red cells is reduced under conditions of anaemia, and given that haemoglobin is found inside red cells, it might be expected that the concentration of HbA1c should fall in people suffering from anaemia. The results also show that the higher the concentration of albumin, the higher that of fructosamine. Serum albumin is the main substrate to which

glucose binds, forming serum fructosamine.

Marginal variances

Gender had no influence on the variability of fructosamine. HbA1c variability, however, was significantly greater in men. See Table 4.7.

The variability plots suggest that variations in HbA1c and fructosamine are greater at higher glucose concentrations, identifying people with diabetes and prediabetes. The wide variability in the glycated proteins at lower glucose concentrations might also identify people with diabetes who are being treated with anti-diabetic drugs, and who have low fasting glucose concentrations (Figure 4.13). The variability of both HbA1c and fructosamine increased with age.

Dependence

Gender had no influence on the association between HbA1c and fructosamine (Table 4.7). As expected, the association between HbA1c and fructosamine is strengthened with increasing fasting plasma glucose, and with age. It is worth noting that Kendall's τ is high for high fasting plasma glucose. In normoglycaemic subjects (plasma glucose < 100 mg/dL), no association was seen between the response variables. See Figure 4.14.

The strength of the association between the glycated proteins was variable, reflecting how the lifespan of the red blood cells may be shortened in people with anaemia.

As a whole, the above findings could have important clinical implications when diagnosing prediabetes - a condition characterized by slightly elevated blood glucose concentrations, and indicative of risk of progression to diabetes - and it should be taken into account that discordances in these markers may be common, especially in young people or in those with prediabetes or early stage diabetes.

One of the advantages of copula regression models is the possibility of deriving further interpretable results from the fitted models. For example, Figure 4.15 shows the contours of the fitted bivariate distribution for different plasma glucose conditions - normo-glycaemic, prediabetes, and diabetes; criteria from American Diabetes Association (2018). It can be seen that the relationship between markers of glycemic control is varying according the levels of plasma glucose values. Correlation is high in patients with diabetes and there is no correlation in normo-glycemics.

Figure 4.16 shows the joint probability of exceeding certain thresholds for some covariates at different ages. It can be displayed that the probability of finding both markers below the diagnostic threshold decreases with age.

Note that, in Figure 4.15 the effect of gender was set to women while all continuous covariates except for glucose were fixed at their mean values for the en-

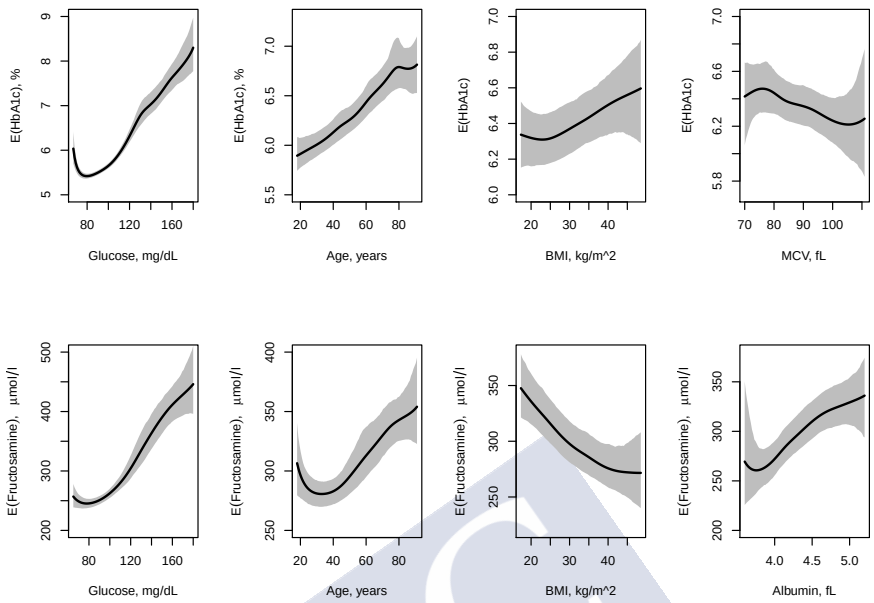


Figure 4.12: Smooth effect of glucose, age, BMI, albumin and MCV on the mean of HbA1c and fructosamine levels.

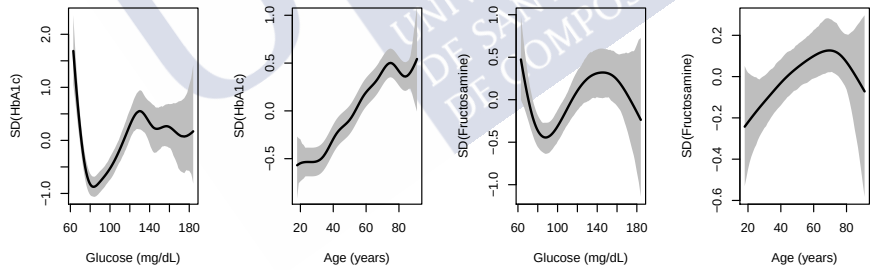


Figure 4.13: Smooth effect of glucose and age on the standard deviation of the HbA1c and fructosamine levels.

tire data set. In Figure 4.16, the effect of gender was also set to women while all continuous covariates except for glucose and age were fixed at their mean values for the entire data set.

In this work, the Gumbel copula provides the best statistical fit (see Section 3.2.3 of the manuscript) and the best clinical explanation. Figure 4.10 shows the dispersion diagram for HbA1c and fructosamine; one can see how this copula captures the relationship between the variables better than either the Gaussian or Clayton copulae (see also Figure 4.1). The biomedical literature reports that the strongest association between glycated proteins is seen when glucose values

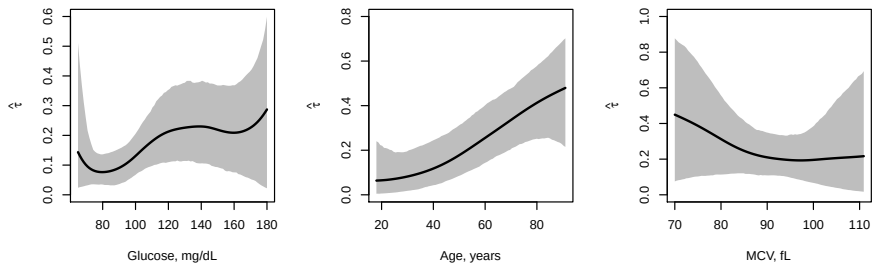


Figure 4.14: Estimates for τ from a Gumbel copula model with log-normal margins for both, HbA1c and fructosamine.

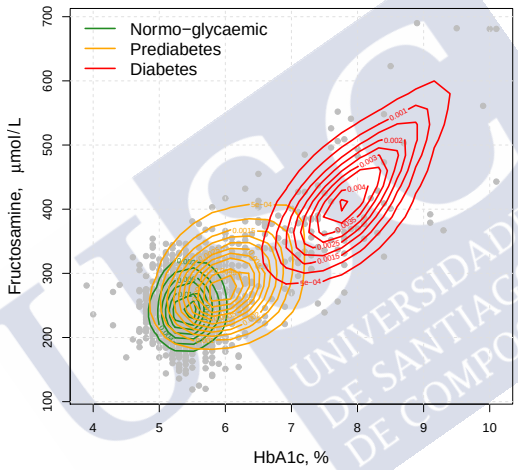


Figure 4.15: Contour lines of densities for three different glucose levels: Normo-glycaemic (FPG $<$ 100 mg/dL or HbA1c $<$ 5.7%); Prediabetes (100 mg/dL $\leq$ FPG $\leq$ 125 mg/dL or 5.7% $\leq$ HbA1c <6.5%); Diabetes (HbA1c $\geq$ 6.5% or FPG $>$ 125 mg/dL) (American Diabetes Association, 2018). In this figure all remaining non-linear effects (except glucose) are kept constant at $f(\overline{\nu})$ (estimated functions evaluated at mean covariate values). Gender has fixed to women.

are higher (Juraschek et al., 2012). Indeed, one of the aims of the chapter was to demonstrate that the relationship between HbA1c and fructosamine changes with the blood glucose concentration (see Figures 4.15 and 4.16). At lower glucose values, no, or only a weak, relationship exists between these variables, while at higher values the relationship is strong. In different biomedical studies in which the correlation between glucose, HbA1c and fructosamine has been examined, a strong correlation has been reported. However, this relationship was usually studied in persons with diabetes, in whom both biomarkers are present at higher concentration. When it is studied in persons who are normo-glycaemic,

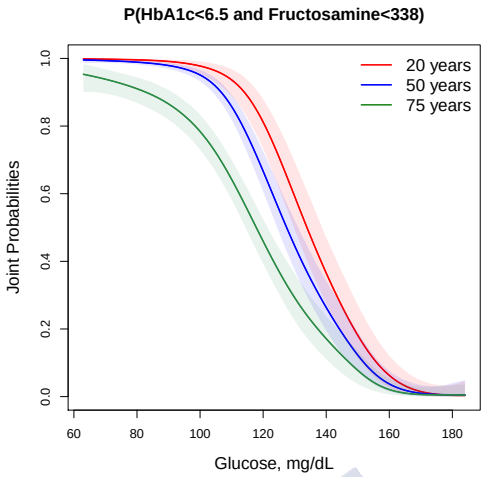


Figure 4.16: Joint probabilities with confidence bands in terms of the glucose values for three different age levels.

in whom the concentrations of both biomarkers are lower, no relationship is seen. This type of dependence - weaker at lower biomarker concentrations - may explain the discrepancies encountered when these biomarkers are used indiscriminately for the diagnosis and monitoring of a particular disease. Therefore, the use of a Gumbel-like copula becomes important.

In studies in which the correlation between the response variables is strong for lower values of a determined covariate, a Clayton-like copula may provide a better fit than a Gumbel copula. In contrast, in those in which there are no tail dependences, the Gaussian copula might be a better option for modelling the relationship between the response variables.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 5

Distributional regression models including functional data

Due to the technological explosion of the last years, data sets in which measurements consist of curves or images instead of scalars (i.e., functional data defined as in Ramsay and Silverman, 2005) are becoming increasingly common in many applied areas (Febrero-Bande and Oviedo de la Fuente, 2012; Scheipl et al., 2016). This new era requires the development of advanced techniques to analyse and draw reliable conclusions from these data. In this context, Figure 5.1 shows an example of a classical functional data such as the considered in the present thesis. But, what is exactly "functional data"? To provide a general definition of this concept, we refer to the definition found in Ferraty and Vieu (2006)

"A random variable $\mathcal{Z}$ is called functional variable if it takes values in an infinite dimensional space (or functional space). An observation Z of $\mathcal{Z}$ is called a functional data."

It should be noted that $\mathcal{Z}$ denotes a random curve, and Z its observation. In the framework of this thesis, we will consider the special case in which the functional data represents a set of curves, i.e., $Z = \{Z(t), t \in \mathcal{T}\}$ where $\mathcal{T}$ is a compact interval such that $\mathcal{T} \subset \mathbb{R}$. [1] From a theoretical point-of-view, this definition assumes that we are dealing with curves measured on arbitrary fine grids, however in practice, the observations are always only taken at a discrete set of time points (Brockhaus et al., 2018). In the remaining of this chapter, we will denote by $Z_i(t), t \in \mathcal{T}, i = \{1, \ldots, n\}$, a functional data taking values in $L^2(\mathcal{T})$[2],

[1] One important thing to notice here is that the concept of functional variable is more general than curve analysis (Ferraty and Vieu, 2006). For example, random surfaces - *as the grey levels of an image or* a *vector of curves* (in which case $\mathcal{T} \subset \mathbb{R}^2$) - are also considered functional variables such as defined in Ferraty and Vieu (2006). In particular, in biomedicine, *neuronal networks* or *medical images in two or higher dimension object data* are examples that currently being studied as functional data; see for example Rossi et al. (2005); Tian (2010); Li et al. (2014); Gruen et al. (2017) or Kendrick et al. (2017).

[2] $L^2(\mathcal{T})$ denotes the space of square integrable functions on the compact real interval $\mathcal{T} =$

and the respectively observed functional data by $z_{ij}(t)$, where $i = \{1, \ldots, n\}$, and $j = \{1, \ldots, R\}$ with observations time points $t \in (t_1, \ldots, t_R) \in \mathcal{T}$. In the above expression, n denotes the total number of individuals and R the total number of time points where the curves are measured. In this work, taking into account the nature of the data considered,we suppose an equally spaced time series, but unevenly spaced time series could also be considered.

As mentioned in the previous chapter, in diabetes fasting plasma glucose (FPG) and glycated haemoglobin (HbA1c) are the main parameters of the glucose metabolism which are used to diagnose and control the hyperglycemia (American Diabetes Association, 2018). More recently, particular emphasis has been given to non-fasting, mainly post-prandial plasma glucose (PPG) as a parameter to be included when assessing metabolic control in patients with diabetes.

The results of interventional studies support the evidence that the reduction of FPG and HbA1c did not achieve a significant decrease in cardiovascular disease, possibly the major complication of type 2 diabetes (U.K. Prospective Diabetes Study (UKPDS) Group, 1998a,b). In these trials the focus was on FPG and HbA1c and not attention was paid to control plasma glucose in the postprandial period, which contribute to HbA1c, but to a lesser extent than fasting glucose concentrations.

The exact contribution of postprandial blood glucose excursions to the overall glycemic control of patients with type 2 diabetes remains largely undetermined (American Diabetes Association, 2018). Some researchers found that postlunch plasma glucose (PG) concentrations were better correlated to HbA1c than fasting values (Bonora et al., 2001); in the same type patients, other reported that preprandial PG concentrations were related to HbA1c more strongly than postprandial concentrations (Monnier and Colette, 2009). Meanwhile, Hashimoto et al. (1995) held that HbA1c mainly reflects mean plasma glucose but it does not reflect glycemic excursion well. Hashimoto et al. (1995) hypothesize that is because a transient increase in plasma glucose reversibly produces unstable HbA1c but does not produce stable HbA1c rapidly. However, the Amadori reaction of glycated albumin (GA) progresses rapidly, unlike HbA1c. GA, another glycemic control indicator, is a glycated protein similar to HbA1c (American Diabetes Association, 2001). They consider that GA reflects glycemic excursion and/or postprandial hyperglycemia in addition to mean plasma glucose (Hashimoto et al., 1995).

The aim of this was to determine the contribution of postprandial plasma glucose on the levels of glycated proteins (HbA1c and fructosamine) by using continuous glucose monitoring in a large number of patients. In this chapter, we assess the effect of postprandial glucose concentrations on the levels of fructosamine, taken the postprandial glucose profiles for three hours after the breakfast as the functional covariate, and the fructosamine as the response in the framework of DR models.

$[T_1, T_2] \subset \mathbb{R}$.

5.1 Introduction

Classical techniques of regression were extended during the last 10 years to the functional context allowing functional data in both, response distribution, and covariates. Functional linear regression models (FLRM) have been increasingly gaining popularity over the years. FLRM can be classified depending on the space where response and covariates take their values: i) scalar response FLRM or scalar-on-functional regression; ii) functional response FLRM; and iii) fully FLRM including both a functional variable response and functional covariates. In this chapter, we will focus on scalar-on-function specifications because of the nature of the presented data and research questions at hand. Specifically, our main objective here is to study how to incorporate functional variables as predictors in DR models.

Regression models that consider a scalar response and a functional covariate are probably one of the most developed functional data areas in recent years (Morris, 2015). Nowadays, the majority of papers have discussed functional predictor regression models based on the idea of FLRM, first introduced by Ramsay and Dalzell (1991). In the literature, various possibilities to estimate functional coefficients in this type of models are available. See, for instance, Morris (2015) for a recent review. Below we mention some of them (Ramsay and Silverman, 2005)

- *Multivariate analysis techniques:* for example i) finite number of *functional principal components* to reduce the dimensionality of the curves for estimating the functional linear model (as for example Cardot et al., 1999; Escabias et al., 2004; Cardot and Sarda, 2005; Hall et al., 2006). ii) *Partial Least Squares*: Preda and Saporta (2005) or Krämer et al. (2008). There are also some works that combine both approaches, see Febrero-Bande et al. (2017) for an overview and a comparative study.

- *The use of fixed basis functions* such as splines or P-splines (Eilers and Marx, 1996), Wavelets or Fourier basis and regularization across $Z_i(t)$, following the general strategy laid out by Ramsay and Silverman (2007) (as Cardot, 2002; Cardot et al., 2003; Antoniadis and Sapatinas, 2003, among others). There are also some papers that use principal components and splines together to perform regularization such as James and Silverman (2005), Yao and Lee (2006) or Reiss and Ogden (2007), among others. Wood (2017) proposed to incorporate functional effects into GAMs using spline expansions.

- *Linear discriminant analysis:* James and Hastie (2001) proposed a regression model including a functional covariate where the curves are irregularly sampled.

- Escabias et al. (2004); James and Silverman (2005); Cardot and Sarda (2005) extended GLM regression models to include a functional covariate and of-

fered some examples including the case of one continuous dependent variable. Müller and Stadtmüller (2005) proposed a new regression technique called *generalized functional linear model*.

In this chapter, we focus on the signal regression model and illustrate how to include functional data as a covariate in DR based on a boosting approach introduced by Brockhaus et al. (2018). The rest of the chapter is organized as follows: Section 5.2 describes the boosting approach of Brockhaus et al. (2018) and proposes a way to include functional data in DR . Section 5.3 provides a simulation study. Lastly, in Section 5.4, this proposal is applied to a clinical case in which fructosamine concentrations are modelled according to postprandial glucose levels, measured over 3 hours after the breakfast as a functional covariate.

In the remainder of the chapter, we will denote the approach of Brockhaus et al. (2018) by *boosting* and our proposal will be denoted by *MCMC*.

5.2 Signal regression effects

The main difference between classical regression and signal regression models lies on the regression coefficients. In signal regression, regression coefficients are functions taking values in time, $\beta(t)$, instead of vectors as in classical regression

$$g(E(Y_i)) = \beta_0 + \int_T \mathbf{Z}_i(t)\beta(t)dt + \epsilon_i, \tag{5.1}$$

where g is a link function, $Y_i, i = 1, \ldots, n$ is a continuous response, β_0 is the overall intercept, $\mathbf{Z}_i(t)$ is a functional predictor and $\beta(t)$ is the functional coefficient to be estimated, and $\epsilon_i \sim N(0, \sigma^2)$ denotes the residual errors.

As mentioned before, the majority of papers in the area of signal regression have focused on modelling only the mean. Recent studies have combined signal regression specifications with GAMLSS (see, Brockhaus et al., 2018; Greven and Scheipl, 2017), whereas Wood (2017) proposed to incorporate functional effects into GAMs using spline expansions. In this chapter, we propose to incorporate functional effects into DR based on the signal regression proposed by Brockhaus et al. (2018). Furthermore, we compare the performance of this new implementation with the proposal of Wood (2011, 2017).

In this section, we discuss how to incorporate functional covariates in DR, i.e, to extend equation (5.1) to accommodate possible functional covariates for each parameter of the response distribution

$$\eta_i^{\vartheta_k} = g_k\left(\vartheta_{ik}\right) = \beta_0^{\vartheta_k} + \int_T \mathbf{Z}_i(t)\beta^{\vartheta_k}(t)dt + \epsilon_i. \tag{5.2}$$

Let us remember that the formulation of a structure additive DR (see Section 3.3 of Chapter 3) is the following

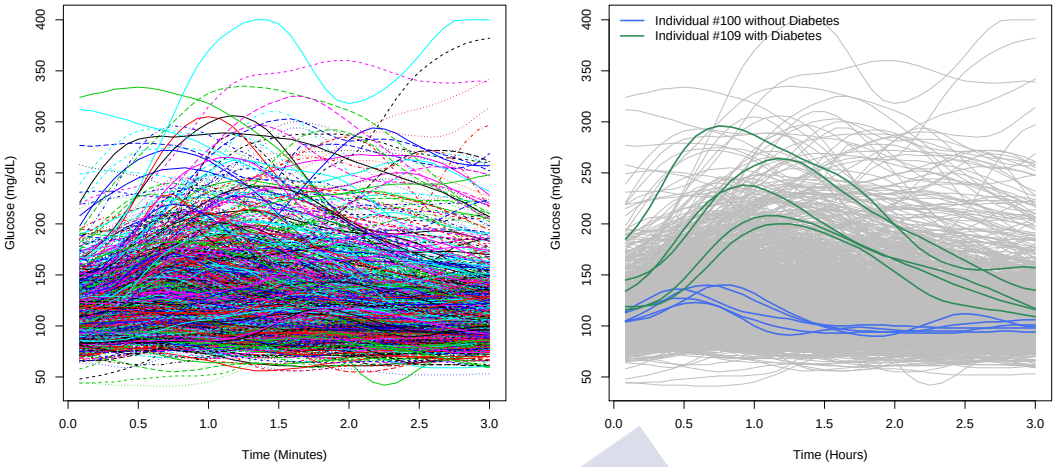


Figure 5.1: The left panel shows glucose profiles of AEGIS participants measured along three hours after breakfast during 5 days. As an illustration, the right panel shows the glucose profiles of two individuals, one with diabetes and the other one without this disease.

$$\eta_i^{\vartheta_k} = \beta_0^{\vartheta_k} + \sum_{j=1}^{J_k} f_j^{\vartheta_k}(\boldsymbol{\nu}_i), j = 1, \ldots, J_k,$$

where f_j represents the different covariate effects. Note that each distribution parameter may depend on different covariates and a different number of effects, say J_k.

The smooth effects are approximated by a linear combination of appropriate basis functions

$$f_j^{\vartheta_k}(\boldsymbol{\nu}_i) = \sum_{d_j=1}^{D_j} \boldsymbol{\beta}_{j,d_j}^{\vartheta_k} B_{j,d_j}^{\vartheta_k}(\boldsymbol{\nu}_i),$$

where $\boldsymbol{\beta}_j^{\vartheta_k}$ is the vector of coefficients (with dimension D_j) to be estimated and $B_{j,d_j}^{\vartheta_k}(\boldsymbol{\nu}_i)$ the basis functions evaluated at the observed covariate values. The superscript $\vartheta_{ik}, k = 1, \ldots, K$ refers to each of the K-distribution parameters of the response variable. In the following, to clarify the notation, we have skipped the subscripts d_j and ϑ_{ik}.

Let be $Z_j(t), t \in T$ one functional covariate observed on $(t_1, \ldots, t_R)'$. The signal regression term, $\int_T Z_j(t)\beta_j(t)$, can be considered in the above framework. More specifically, Wood (2011) propose to approximate the integral $\int_T Z_j(t)\beta_j(t)dt$ using integration weights (Δ) as follows

$$\begin{aligned}\boldsymbol{B}_j(Z_j(t))' &= \left[\tilde{Z}_j(t_1), \ldots, \tilde{Z}_j(t_R)\right] [\boldsymbol{b}_j(t_1) \ldots \boldsymbol{b}_j(t_R)]' = \\ &= \left[\sum_{r=1}^{R} \tilde{Z}_j(t_r)\boldsymbol{b}_1(t_r) \ldots \sum_{r=1}^{R} \tilde{Z}_j(t_r)\boldsymbol{b}_{Q_j}(t_r)\right].\end{aligned} \tag{5.3}$$

In this expression, $\tilde{Z}_j(t)$ is defined as $\Delta(t)Z_j(t)$ and $\boldsymbol{B}_j(t)$ is a vector of basis functions evaluated at t, $\{\boldsymbol{b}_q(t), q = 1, \ldots, Q_j\}$. Choices of the basis functions and adequate penalties matrix ($\boldsymbol{P}_j$) depend on the investigator, for example, in this framework Bayesian P-splines (Eilers and Marx, 1996; Fahrmeir and Kneib, 2011) with second order differences will be considered. However another basis functions could be considered such as thin plate regression splines (Wood, 2011), or Functional Principal Component (FPC, Ramsay and Silverman, 2005) basis.

In line of Brockhaus et al. (2018), we propose to extend the gradient boosting for GAMLSS with functional covariates to the framework of DR. Before that, Section 5.2.1 summarizes the most relevant aspects of this boosting methodology.

5.2.1 Boosting approach

Gradient boosting is a general method with the main objective of optimizing an expected loss criterion along the steepest gradient boosting (Freund and Schapire, 1996), i.e.

$$\hat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}} \mathbb{E}\left[\varphi\left(Y, \boldsymbol{f}(X)\right)\right], \tag{5.4}$$

where $\boldsymbol{f}(X) = (f_1(X), \ldots, f_K(X))'$ and $\varphi : \mathbb{R} \times \mathbb{R}^K \to \mathbb{R}^+$ is the loss function.

Let us assume that observations $(y_i, \nu_i, i = 1, \ldots, n)$ are made, equation (5.4) can be approximated by

$$\hat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}} \frac{1}{n} \sum_{i=1}^{n} \left[\varphi\left(y_i, \boldsymbol{f}(\nu_i)\right)\right]. \tag{5.5}$$

Brockhaus et al. (2018) have extended the gradient boosting algorithm proposed in Mayr et al. (2012a) for GAM regression models (see Hofner et al. (2014) for a description of the available R-software and a theoretical background of this methodology) to GAMLSS framework.

The gradient boosting algorithm permits to represent the model as the sum of simple penalized regression models, called base-learners for fitting GLM and GAM - to potentially high-dimensional data - in the case of Hofner et al. (2014) and for GAMLSS in Brockhaus et al. (2018).

To estimate a GAMLSS via boosting, Brockhaus et al. (2018) have proposed to use a component-wise gradient boosting algorithm (Breiman, 1998; Friedman, 2001). In the following, we summarized the algorithm proposed by Brockhaus et al. (2018)

(i) The loss function is defined as the negative log-likelihood of the response distribution as follows

$$\varphi(y_i, \boldsymbol{f}(\nu_i)) = -l\left(y_i, \boldsymbol{\vartheta}_i\right),$$

note that the log-likelihood, denoted by l depending on the response, y_i and each parameter of the response distribution, i.e, $\boldsymbol{\vartheta}_i = (\vartheta_{i1}, \ldots, \vartheta_{iK})$, with $\vartheta_{ik} = \left(g^{\vartheta_k}\right)^{-1}\left[f_j^{\vartheta_{ik}}(\nu_i)\right]$. Let us recall that, g is the corresponding link function for each distribution parameter ϑ_{ik}.

(ii) The user must i) define the bases, $B_{j,d_j}^{\vartheta_k}$ and the desired penalties; ii) select a vector of step-lengths $(v^{(1)}, \ldots, v^{(K)})$ whose components must be in the interval $(0, 1)$; iii) a vector of stopping iterations $(m_{\text{stop}}^{(1)}, \ldots, m_{\text{stop}}^{(K)})'$; iv) initialize the coefficients $\boldsymbol{\beta}_j^{\vartheta_k[0]}$; and finally v) set the number of boosting iterations to zero, i.e, $m^{(k)} = 0$.

(iii) The component-wise gradient boosting is a machine learning for optimization and obtain estimated models using gradient descent techniques with a similar interpretation that as classical regression models (Hofner et al., 2014). The algorithm will be iterate over each parameter of the response distribution, i.e, $\vartheta_k, k = \{1, \ldots, K\}$ following the next steps

(a) Set $k = 1$.

(b) If the number of boosting iterations, $m^{(k)}$, is higher than the stop iteration, i.e, $m_{\text{stop}}^{(k)}$, go to *step (iii) point (g)*; else, compute the negative partial gradient of the loss function $\left(-\frac{\partial}{\partial \boldsymbol{f}}\varphi(y, \boldsymbol{f})\right)$ - by plugging in the current estimates $\hat{\boldsymbol{f}}^{[m]} = \left(\hat{f}^{(1)[m]}, \ldots, \hat{f}^{(K)[m]}\right)'$, where $\hat{f}^{(k)[m]}(\nu_i) = \sum_j \boldsymbol{\beta}_j^{\vartheta_k[m]} \boldsymbol{B}_j^{\vartheta_k}(\nu_i))$ - and evaluate at $\hat{\boldsymbol{f}}^{[m]}(\nu_i)$ as follows

$$u_i^{(k)} = -\frac{\partial}{\partial f^{(k)}}\varphi(y_i, \boldsymbol{f}(\nu_i))\Big|_{\boldsymbol{f}(\nu_i)=\hat{\boldsymbol{f}}^{[m]}(\nu_i).}$$

(c) Fit the base-learners contained in $f^{(k)}$ for $j = 1, \ldots, J_k$ to $u_i^{(k)}$ taking into account the appropriate penalties.

$$\widehat{\gamma}_j = \underset{\gamma \in R^{Q_j}}{arg\,min} \sum_{i=1}^{n}\left[u_i^{(k)} - \boldsymbol{B}_j^{\vartheta_k}(\nu_i)'\boldsymbol{\gamma}\right]^2 + \boldsymbol{\gamma}'\boldsymbol{P}_j^{\vartheta_k}(\boldsymbol{\lambda})\boldsymbol{\gamma}.$$

(d) Select the base-learner which provides the best fit according to least squares criterion

(e) Update the $f^{[m](k)}$ coefficients to $\boldsymbol{\beta}_{j^*}^{\vartheta_k[m]} = \boldsymbol{\beta}_{j^*}^{\vartheta_k[m]} + v^{(k)}\widehat{\gamma}_{j^*}$.

(f) Set the coefficient $\boldsymbol{\beta}_j^{\vartheta_k[m+1]} = \boldsymbol{\beta}_j^{(k)[m]}$, $j = 1, \ldots J^{(k)}$, for those that do not correspond to the selected base-learner.

(g) If $k \neq K$ - remember that K denotes the total number of the parameters of the response distribution - increase k by 1 and go back to iii) point (b).

(iv) Iterate Step iii) until the stopping iteration $m_{\text{stop}}^{(k)}$ is reached. See Mayr et al. (2012b) and Brockhaus et al. (2018) for a discussion of the importance of knowing when to stop and how to select the number of stopping iterations.

It should be noted that each component of the final model is a linear combination of each base-learner fit (Brockhaus et al., 2018) and the number of boosting iterations are the only tuning parameters. Note that the user can use a different number of boosting iterations for each distribution parameter.

In R software, `FDboostLSS` function of `FDboost`-package[3] fits GAMLSS regression models via component-wise boosting. This function, allows to include functional covariates or even a functional response into a GAMLSS regression model. This package is a natural extension to the ones commonly known `mboost` package - which allows to estimate, GAM and GLM models via component-wise boosting - to the `GAMLSS` package.

5.2.2 MCMC approach

The approach defined by Brockhaus et al. (2018) can immediately be cast into the DR framework by replacing the usual design matrix and penalty matrix by versions that originate from functional data. As a consequence, no changes in MCMC Bayesian algorithms are required. We can refer to DR Bayesian algorithms discussed in Chapter 2 and 3.

To estimate the proposed model, the BayesX software (Belitz et al., 2016) can be used. Appendix C includes a guide on how to use these models, in practice. It should be noted that functional terms can be included in all the equation of the DR model - as showed in Appendix C. Furthermore, these terms can be easily mixed with other types of smooths as non-linear, spatial-temporal or random effects, for instance. The different terms have to be separated by "+" signs in the desired equations as showed in Appendix A.

5.3 Performance of the proposed MCMC model

In this section, we will study the performance of the proposed *MCMC* method. Furthermore, we will compare the results obtained with the linear functional ef-

[3] `FDboost`– package fits regression models for functional data, scalar-on-function, function-on-scalar and function-on-function regression models by a component-wise gradient boosting algorithm. See Brockhaus et al. (2017); Brockhaus and Ruegamer (2018) for more details.

fects proposed by Wood (2017) and Brockhaus et al. (2018). To the best of our knowledge, both approaches are the ones available on the statistical literature that allow to model each parameter of the response distribution taking into account the effect of functional data among other covariates. There follows an introduction about the frequentist approach proposed by Wood (2017) and then the simulation study will be presented.

5.3.1 Frequentist approach

Because of the very modular structure of the algorithm underlying the GAM (Wood, 2003) and GAMLSS (Rigby and Stasinopoulos, 2005) approach, the scope of these models could be straightforwardly extended in the direction of signal regression.

Wood (2017) proposes a signal regression approach to account for this type of covariate effect by using

$$\int s(t)Z_i(t)dt, \tag{5.6}$$

where s is a smooth function to be estimated. Since integration is a linear operation, such effect can be estimated using the same smoother as continuous covariates (described in Section 2.1.1 of Chapter 2) but with a `by` factor. In particular, Wood (2017) proposes replacing (5.6) with a discrete sum approximation that allows to estimate the functional covariate using the summation convention that is employed when a smooth term is supplied with matrix arguments.

The main differences between boosting algorithm and the common estimation of the penalized likelihood is that in boosting each effect is estimated separately as a base-learner. However in penalized likelihood estimation all effects are estimated at the same stage.

In practice, we will use the function `s()` of GJRM R-package (Marra and Radice, 2017a) [4]. This package relies on Simon Wood's methodology (Wood, 2017). To illustrate this, let `Z` be the $n \times J_R$ matrix, with one profile per row, and `Time` be the matrix of times at which the functional values are measured (all rows are identical). Note that `Time` and `Z` are matrix with identical dimensions. Then, the smooth formula one has to use to adjust for such functional effect is `s(Time, by = Z)`. As in the *MCMC* approach functional terms can be included in all the equations and these terms can be combined with other types of covariates. In the remaining of the chapter, this approach will be denoted by *frequentist*.

[4] We propose the use of GJRM package (instead of `mgcv`) because its additive predictors' set up depends on the `mgcv`-package (Wood, 2017) and the current implementation supports 18 continuous and discrete distributions. The current version of `mgcv`-package allows only a continuous response and two categorical responses in the framework of GAMLSS.

5.3.2 Simulation study

The aim of this simulation study is to check the performance of combining boosting techniques as proposed by Brockhaus et al. (2018) with DR (as defined in Klein et al., 2015) (approached denoted by *MCMC*). Furthermore, we will compare the results obtained with the *frequentist* and *boosting* approaches presented below. The considered methodologies were tested for different scenarios with scalar covariates. For this reason, this simulation study is focused on the estimation of functional data covariates.

Data generation

For the generation of the data, we choose a similar pattern as that in Brockhaus et al. (2018). Specifically, in this study, we consider $R = 1000$ replications of sample sizes $n = 500$ and $n = 1000$ for the following settings.

First, we consider a normal distributed response (y_i) where, both the expectation and standard deviation, depend on one functional covariate $Z_i(t), i = 1 \dots n, t \in [1, 2]$ as follows

$$\begin{cases} \eta_i^{\mu} = \beta_0^{\mu} + \int z_{ij}(t)\beta_j^{\mu}(t)dt \\ \eta_i^{\sigma} = \log(\sigma) = \beta_0^{\sigma} + \int z_{ij}(t)\beta_j^{\sigma}(t)dt. \end{cases} \tag{5.7}$$

The first equation of formula (5.7) refers to the location parameter while second equation refers to the scale parameter. Hereafter the parameter index is eliminated for the sake of simplicity. The predictors (η_i) are formed through the additive composition of an intercept β_0 representing the overall level of the predictor, and coefficients $\beta(t)$ reflecting the effect of the functional covariate.

- Functional covariates are generated with 40 equally spaced evaluated points $(t_1, \dots, t_{40})'$, using the following 5 basis functions:

  $$\phi_c(t) \sin((c - 0.5)t)), c = 1, \dots, 5$$

  with random coefficients from a C-dimensional normal with $N_C(\mathbf{0}, \boldsymbol{I})$.

- Furthermore, for the simulation study, the covariates were centred by subtracting the column means.

- The coefficient functions to model the expectation $(\beta_j^{\mu}(t))$ and the standard deviation $(\beta_j^{\sigma}(t))$ - for the functional covariate - have been generated as follows

  - $\beta_j^{\mu}(t) = \sin(1.75\pi t - 1)$
  - $\beta_j^{\sigma}(t) = \cos(1.4\pi t - 1.3\pi)$.

Estimation details and results obtained

- For the estimation of the models, normal location scale models have been considered in the three approaches considered.
- Functional covariates have been estimated using Bayesian P-splines (Eilers and Marx, 1996; Fahrmeir and Kneib, 2011) with second order difference penalties.
- For the estimation by boosting, we have fixed 300 stop boosting iterations for β_j^μ and β_j^σ and the step-length considered was $v = (0.1, 0.1)$.
- To evaluate the performance of considered approaches, mean-squares error (MSE) has been computed over the domain of the functional covariate for each approach

$$MSE(\beta_j^\mu(t)) = \int \left\{ \beta_j^\mu(t) - \hat{\beta}_j^\mu(t) \right\}^2 dt,$$

and

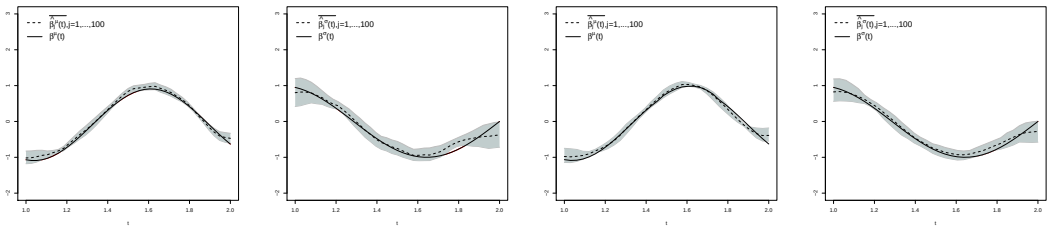
$$MSE(\beta_j^\sigma(t)) = \int \left\{ \beta_j^\sigma(t) - \hat{\beta}_j^\sigma(t) \right\}^2 dt.$$

Results obtained are summarized in Figures 5.2, 5.3, 5.4 and 5.5. Figure 5.5 shows the MSE's obtained with the three approaches considered: *frequentist*, *MCMC* and *boosting*. As expected, results are better at higher sample size. In the considered scenario, both *MCMC* and *boosting* approaches, capture the maximum of the function better than *frequentist* approach, (see Figures 5.2 and 5.3). Figures from 5.2 to 5.4 also shows that estimated coefficient functions are close to the true functions in all the approaches. Observed differences in mean of MSE between the three approaches are less than 0.012 in all the cases considered here. $MSE(\beta_j^\mu(t))$ (resp. $MSE(\beta_j^\sigma(t))$) differences between *boosting* and *MCMC* are less than 0.003 (resp. 0.009) - in mean.

It should also be noted that MSE's values of mean coefficient functions are smaller than standard deviation, see Figure 5.5. This has already been detected in Brockhaus et al. (2018).

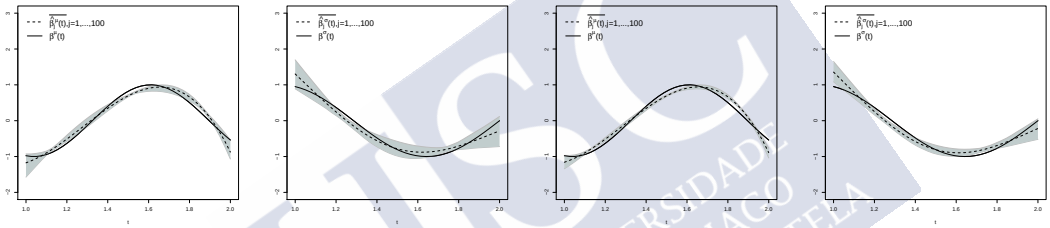
5.4 Continuous glucose monitoring: Application to AEGIS

The proposed MCMC approach will be used in this section to undertake a specific study of fructosamine concentrations in relation to glucose profiles and other clinical covariates in an adult population-based survey. In the following, we describe the data, the model building process, and comment on the results obtained.



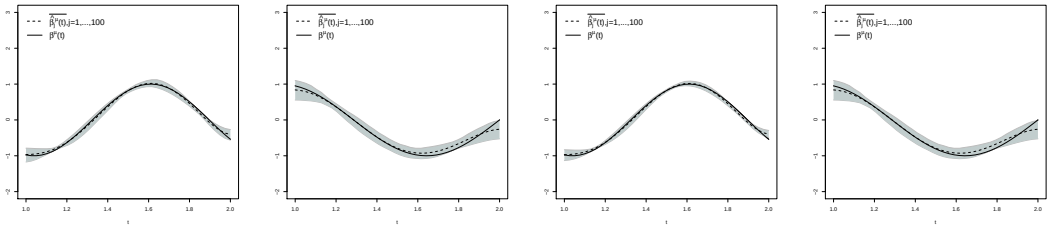
a) n=500 ***b) n=1000***

Figure 5.2: Functional coefficients estimates for mean (first and third panel) and sigma (second and fourth) obtained by using the *MCMC* approach to data simulated. True coefficients are represented by black solid lines, mean estimates by dashed lines and pointwise ranges resulting 5% and 95% credible interval from the 1000 replications by shaded areas.



a) n=500 ***b) n=1000***

Figure 5.3: Functional coefficients estimates for mean (first and third panel) and sigma (second and fourth) obtained by using the *frequentist* approach to data simulated. True coefficients are represented by black solid lines, mean estimates by dashed lines and pointwise ranges resulting 5% and 95% quantiles from the 1000 replications by shaded areas.



a) n=500 ***b) n=1000***

Figure 5.4: Functional coefficients estimates for mean (first and third panel) and sigma (second and fourth) obtained by using the *boosting* approach to data simulated. True coefficients are represented by black solid lines, mean estimates by dashed lines and pointwise ranges resulting 5% and 95% quantiles from the 1000 replications by shaded areas.

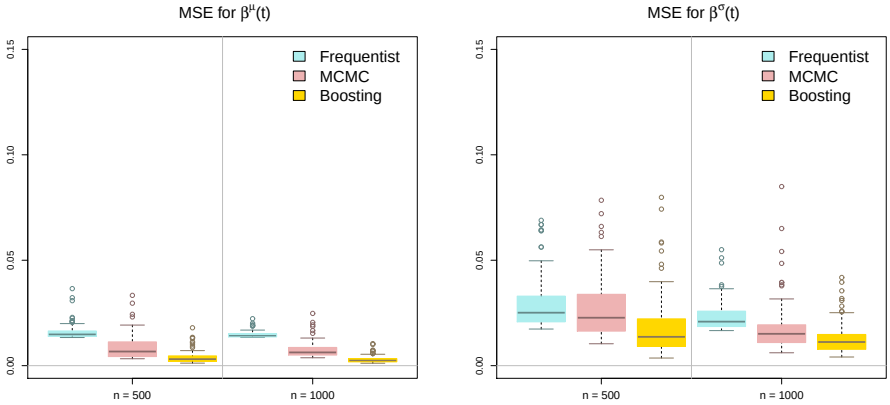


Figure 5.5: MSE of the estimated coefficient functions, $\beta^{\mu}(t)$ (left panel) and $\beta^{\sigma}(t)$ (right panel) for *frequentist* (blue), *MCMC* (pink) and *boosting* (yellow).

5.4.1 Data description

The data - used in this chapter - comes from the AEGIS study, already presented in Chapter 4. A sub-sample of the AEGIS participants were also invited to take part in a project, which included continuous glucose monitoring (CGM) procedures. A total of 622 subjects consented to undergo a 6-day period of CGM. Individuals showing signs of allergy to adhesives or any concomitant medical condition that would likely affect the device evaluation of the performance were excluded.

Before starting CGM, a nurse belonging to the AEGIS team explained to participants the use of the monitoring device, which is carried subcutaneously on the abdomen. This device continuously measures interstitial glucose levels, storing values at five minute intervals. Each participant was also provided with a conventional glucometer, compatible lancets and test strips for calibrating the device. When the monitoring time was complete, the sensor was removed and the data downloaded. When data skipping exceeded 2 hours per day, the data for the entire day were excluded. Of the 622 participants enrolled, 41 were lost to analysis due to difficulties in operating the device or for not adhering to the protocol (Gude et al., 2017).

The values recorded for the first 3 hours following breakfast were those examined in the present work. After excluding, people without diabetes, the final number of samples used in the analysis was 504. Figure 5.1 shows the glucose profiles of these individuals. Given that there are five monitoring valid days per individual, in order to take in account the information available in all the glucose profiles, we propose to use the concept of depth (López-Pintado and Romo, 2009) for functional data profiles (see Figure 5.6).

Different depth measures have been proposed in the statistical literature. For example, for a one-dimensional continuous covariate, the median is usually used

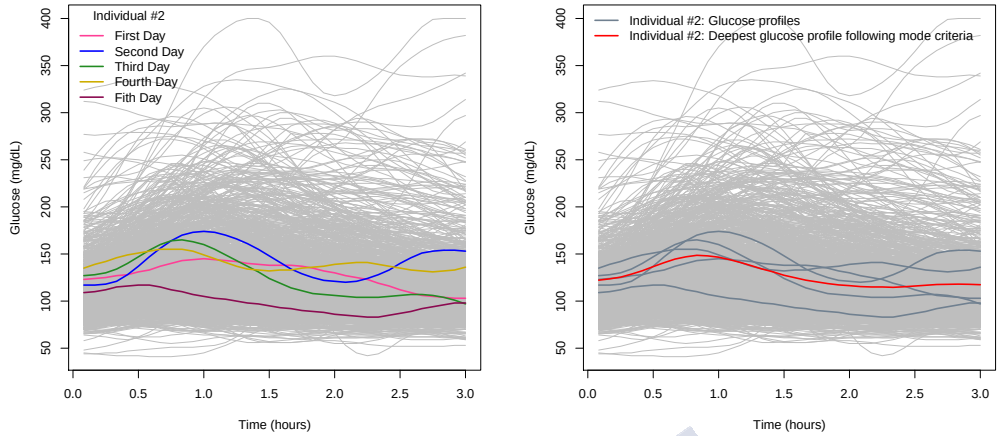


Figure 5.6: Left panel shows the available glucose profiles of a random participant of AEGIS's continuous glucose monitoring measured along 3 hours after the breakfast. Right panel shows the average from the 75% deepest glucose profiles following mode criteria of the random participant selected (in red).

as the deepest point of a cloud of points. But, in functional data, there are more depth notions (Febrero-Bande and Oviedo de la Fuente, 2012). In this application, we consider the h-mode depth, as defined in Cuevas et al. (2007). The population h-mode depth of a functional datum z is defined as the following function (Cuevas et al., 2007)

$$D_h(z) = \mathbb{E}\left(K_h\|z - Z\|\right),$$

where Z is the random element describing the population, $\|.\|$ denotes L^2-norm, h is a fixed tunning parameter, and K_h denotes a re-scaled Kernel, i.e., $K_h(t) = \frac{1}{h}K(t/h)$). In this case, we consider the Gaussian Kernel. Hence, in this biomedical study, we propose to calculate the 0.25-trimmed mean based on h-mode depth (i.e., the average from the 75% deepest glucose profiles following mode criteria) using `fda.usc`-package (Febrero-Bande and Oviedo de la Fuente, 2012). [5]

Let $\text{Glu}_{ij}(t), i = 1, \ldots, 570, j = 1, \ldots, 36, t \in [0, 3]$ be the trimmed mean combining with $h-$mode depth of the glucose profiles. A natural functional descriptive analysis for these glucose profiles is showed in Figure 5.7.

The main aim of this section is to study and quantify the contribution of postprandial glucose profiles on the fructosamine levels, after adjusting for potential confounding covariates. In this context, the following variables were considered to be covariates in order to determine which factors influence in fructosamine

[5] Cuevas et al. (2007) show the good performance of the trimmed mean combining with h-mode depth. López-Pintado and Romo (2009) also show via simulation studies that trimmed mean have better performance than other possible location estimators proposed in the functional data literature.

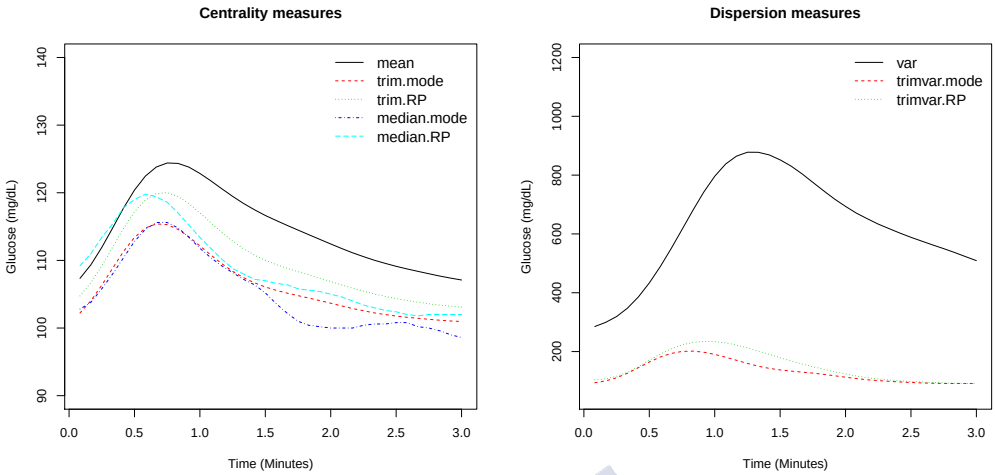


Figure 5.7: Descriptive statistics based on depth of the glucose profiles obtained with `fda.usc`-package (Febrero-Bande and Oviedo de la Fuente, 2012). The left panel displays the i) the `Mean` glucose profiles (defined as, $\overline{Z}(t) = N^{-1}\sum_{i=1}^{N} Z_i(t)$, i.e, the average of the functions point-wise across replications); ii) the average from the 75% deepest glucose profiles following mode criteria (`trim.mode`); iii) the average from the 75% deepest glucose profiles following mode criteria (`median.mode`); and following iv) random projection criteria (`trim.RP`); v) the deepest curve following mode criteria (`median.mode`), and vi) the random projection criteria (`median.RP`) as defined in Cuevas et al. (2007). The right panel gives i) the `variance` glucose profile, (defined as $(N-1)^{-1}\sum_{i=1}\left[Z_i(t)-\overline{Z}(t)\right]^2$); ii) the marginal variance from the deepest curves following mode criteria (`trimvar.mode`); and iii) following the random projection criteria (`trimvar.RP`) (Febrero-Bande and Oviedo de la Fuente, 2012).

levels: age (in years), BMI, (in kg/m^2), albumin levels in f/L [alb in formulae (5.8)] and glucose profiles mg/dL. The scaled fructosamine was considered as the scalar response.

5.4.2 Model building

The response variable, y = fru, defined as fru $= \frac{\text{fructosamine}-\overline{\text{fructosamine}}}{\text{sd(fructosamine)}}$ represents the scaled variable of fructosamine levels.

Assuming a normal as possible response distribution, we propose the following model to study the glycation factors. The two equations of the following formula (5.8) refer to the μ and σ parameters of the response variable y_i with $i = 1, \ldots, n$ and $j = 1, \ldots 36$.

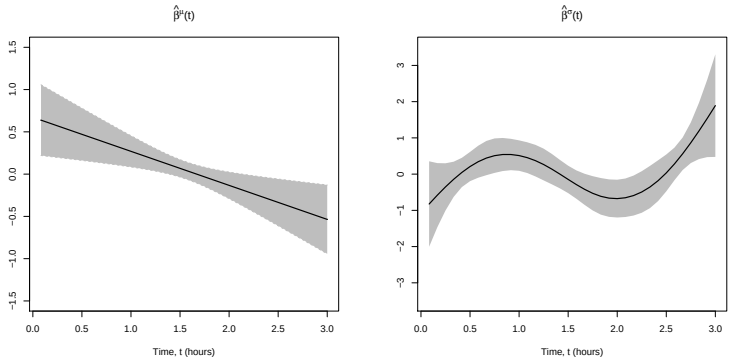


Figure 5.8: Estimated functional glucose coefficients for μ_i and σ_i from model (5.8) using BayesX.

$$\begin{cases} \mu_i = \beta_0^{\mu} + \int \text{Glu}_{ij}(t)\beta_j^{\mu}(t)dt + f_1^{\mu}(\text{Age}_i) + f_2^{\mu}(\text{Alb}_i) + f_3^{\mu}(\text{BMI}_i) \\ \log \sigma_i = \beta_0^{\sigma} + \int \text{Glu}_{ij}(t)\beta_j^{\sigma}(t)dt + f_1^{\sigma}(\text{Age}_i), \end{cases} \tag{5.8}$$

where β_0^{μ} and β_0^{σ} represents the intercept for mean and standard deviation equations, respectively. Eliminating the parameter index for the sake of simplicity, $\int \text{Glu}_{ij}(t)\beta_j(t)dt$ represents the functional effect ot the glucose profiles.

The non-linear effects of continuous covariates (age, BMI, alb) were modelled using Bayesian versions of penalized splines (P-splines, Lang and Brezger, 2004), introduced into a frequentist setting by Eilers and Marx (1996). To model age, BMI, and alb, 20 inner knots, a cubic spline basis, and a second order random walk prior for penalised splines were contemplated. The effects of functional covariates $\text{Glu}_{ij}(t)$ have been estimated as presented in Section using also 20 cubic B-splines basis, and a second order random walk prior for penalised splines.

Results are summarized in Figures 5.8 and 5.9. A significant effect has been detected between glucose profiles, fructosamine and age, BMI, and albumin. It should be understood that the results obtained in Chapter 4 - for the continuous variables in the fructosamine marginal - are very similar here.

Model checking

Quantile residuals - presented in Chapter 3 - can be used to check the performance of model (5.8). Figure 5.10 indicates that our model fits well in almost all the observations. As mentioned before, a normal distribution has been assumed for the response variable in equation (5.8). It can be shown that the results are quite similar with another distribution as the log-normal distribution (data not shown). However, taking into account the good performance presented in Figure 5.10, the Gaussian presumption seems to be adequate taking into account the

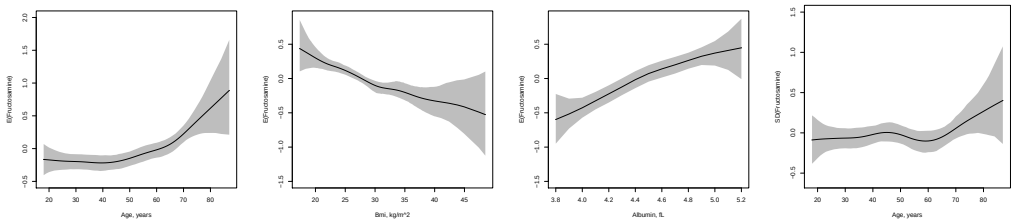


Figure 5.9: The three first plots shows posterior mean estimates of non linear effects of age, BMI, and albumin on the mean of fructosamine. The last one represents posterior mean estimates of non linear effect of age on the standard deviation of fructosamine.

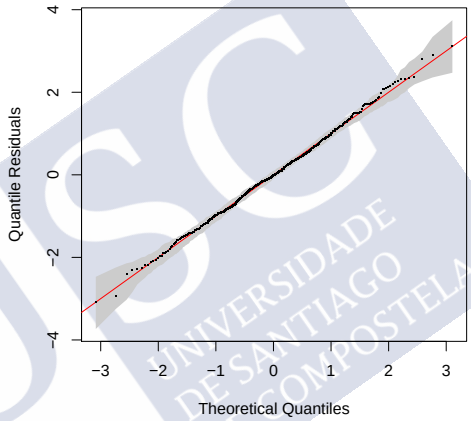


Figure 5.10: Quantile-quantile plot of normalized quantile residuals for model (5.8) with reference bands: the closer the residuals to the bisecting red line, the better the fit to the data.

complexity of the considered data and the sample size. Nevertheless, other complex response distribution could be considered in view of the flexibility which DR regression models allow for response distributions. See Chapters 2 and 3.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 6

Functional regression CGAMLSS

This chapter proposes - for the first time - the incorporation of functional data covariates into copula regression models for location, scale and shape (CGAMLSS). Furthermore, this chapter also propose the use of CGAMLSS with flexible additive predictors (including functional effects) to model the joint distribution of glycated haemoglobin and fructosamine (two proteins that are useful in the control of individuals with diabetes). The level of glucose in patients is an important predictor of these two diabetic proteins, and in our study this is recorded every 5 minutes over several days. Therefore, glucose needs to enter the model as a functional covariate. The inclusion of glucose profiles into this type of models marks a novel contribution in diabetes research.The modelling framework is also extended to include some newly functions that aid the interpretation of the empirical results.

6.1 Introduction

The usefulness of HbA1c and fructosamine proteins in the diagnosis and control of diabetes is based on a relevant phenomenon called "glycation," a process by which glucose is chemically bound to amino groups of proteins. However, the exact contribution of postprandial blood glucose excursions to the overall glycemic control remains largely undetermined. Thus, clinical researchers are interested in the effect of postprandial glucose excursions on the levels and variability of glycated proteins, knowing that both responses are highly correlated depending on the levels of plasma glucose (see Figure 6.1).

Continuous monitoring of subcutaneous glucose values taken every 5 minutes over several days provides a detailed picture of glucose variability. Such profiles have been shown to correlate well with blood glucose levels, although there is a lag between a change in blood value (e.g., after food intake) and the response of the subcutaneous value. This calls for advanced statistical techniques in the fields of multivariate response regression and functional data. Although

the use of standard approaches can go some way to help understand the possible relationships between various factors and a given response, they do not have the necessary in-depth capability to describe the complexity of the problem at hand. Specifically, in classical regression models, it is common to study the mean of the response as a function of some explanatory variables. However, focusing solely on means may lead to an over-simplified picture of the situation. In fact, in many applications, it is important to characterise the effects of covariates on all the parameters of the response's distribution. It is also often necessary to model jointly two or more responses. In many cases, the choice of response distributions is often driven by mathematical convenience. Moreover, flexible covariate effects are not typically considered.

In recent years, various copula-based regression approaches have been developed to model simultaneously two or more responses in the presence of covariates. In this chapter, we have adopted copula generalized additive models for location, scale and shape (CGAMLSS, Marra and Radice, 2017a). In the statistical literature, other copula-based regression approaches have been developed but, as compared to Marra and Radice (2017a), they either only cover some of the flexibility of the CGAMLSS or are based on less efficient estimation approaches (e.g., Yee, 2015; Vatter and Chavez-Demoulin, 2015; Sabeti et al., 2014; Acar et al., 2013; Gijbels et al., 2011). See, Chapter 4 for a recent review of this type of copula regression models.

Klein and Kneib (2016b) offered a software implementation, hence their method seems to be the only readily available competitor to CGAMLSS. In this work, we have adopted the modelling framework of Marra and Radice (2017a) since it allows for more marginal distributions and copulae, and for more general predictor specifications. [1] Furthermore, the CGAMLSS can be easily fitted using the `gjrm()` function in the `R` package `GJRM` (Marra and Radice, 2017b). In fact, it should be mentioned that the availability of software for fitting copula regression models with additive predictors is scarce.

Copula regression techniques may be very useful for clinicians, however studies in which the usefulness of this methodology was explored in the setting of biomedicine are rare and even more so when considering flexible additive predictors. In this study, the CGAMLSS framework has been considered in a biomedical real study about diabetes. For this work, `GJRM` has been extended to include functions that make the output of such models interpretable, an aspect which can encourage the wider uptake of the methods by biomedical researchers but not only. Another important aspect is that of incorporating functional effects (Ramsay and Silverman, 2007) of glucose profiles into the model, something that has not been considered before in the context of CGAMLSS.

Because of the very modular structure of the algorithm underlying the adopted CGAMLSS approach and because its additive predictors' set up depends on the

[1] Note that Chapter 4 shows that, both, the frequentist (Marra and Radice, 2017a) and the Bayesian frameworks (Klein and Kneib, 2016a) of CGAMLSS provide similar results.

`mgcv R` package (Wood, 2017), the scope of CGAMLSS could be straightforwardly extended in the direction of functional regression. As mentioned earlier, this extension was motivated by the interest in modelling the relationship between patients' post-meal glucose profiles and two measures of glycemic control (HbA1c and fructosamine) in a population-based study. With this aim, the AEGIS project was launched and glucose profiles were recorded by means of continuous monitoring devices, in a sizable population sample over the course of one week. See also Chapter 5.

6.1.1 Functional regression effects

Functional data have been extensively studied in the univariate regression context; see, for instance, Morris (2015) and Chapter 5 for a recent review.

In this chapter, we have focused on *functional predictor regression* or *scalar-on-function regression* where the response variable is scalar and there are (at least) one functional covariate because of the nature of the data and research questions at hand.

Functional regression models have been increasingly gaining popularity over the years. In line with scalar-on-functional regression, a lot of research has focussed on functional covariates in regression models. Nowadays, the majority of papers have discussed functional predictor regression models based on the idea of functional linear regression models (FLRM), first introduced by Ramsay and Dalzell (1991). See Chapter 5 for a review of FLRM. Most of these approaches assume a functional linear effect on the response for each unit of time. The principal advantage of this assumption is that they are easy to interpret. However, these type of FLRM are not general enough and can lead to errors when modelling data encompassing to complex structures. To overcome this issue, McLean et al. (2014) proposed the use of bivariate tensor products of B-splines to estimate the functional covariates in a scalar regression model. Alternative approaches to McLean et al. (2014) have been formulated in the statistical literature considering flexible additive structures (Morris, 2015). More specifically, James and Silverman (2005) proposed an approach called functional adaptive model (FAME) which extends GAM models, and projection pursuit regression (Friedman and Stuetzle, 1981) to include functional covariates. On the other hand, Müller and Yao (2008) proposed to use an additive structure. However, James and Silverman (2005) and Müller and Yao (2008), truncate the number of functional principal components to a finite number based on the relationship between predictor and response thorough linear functions of the curves.

Finally, the majority of papers in the area have focused on modelling only the mean of a univariate response (e.g., James and Silverman, 2005; Müller and Yao, 2008; McLean et al., 2014) - as mentioned in Chapter 5. Recent studies have combined linear functional regression specifications with GAMLSS (Brockhaus et al., 2018; Greven and Scheipl, 2017) but only for univariate responses. See Chapter 5

for a review. Compared to previous studies, the adopted CGAMLSS framework allows us to include non-linear functional covariates in a copula model, hence giving rise to a multivariate functional regression approach. This framework allowed us to model flexibly the parameters of the marginal distributions and of the copula parameter (which describes the dependence between the responses), and to combine functional and other types of covariate effects (e.g., non-linear, spatial-temporal or random effects). To the best of our knowledge, only the work by Gijbels et al. (2012) has considered the inclusion of functional predictors in copula regression models. This approach extended the methodology introduced in Gijbels et al. (2011) to estimate the dependence of two response variables conditioned on a functional covariate. However, some of the limitations are that: the authors only considered the effect of a continuous or functional covariate on the dependence, hence not mentioning the possibility of using several types of effects simultaneously; the estimation approach may not efficient since it is based on a two-step procedure; there is no software implementation available.

The remainder of the chapter is organised as follows. Section 6.2 provides an introduction to CGAMLSS, including some details on the smoother set up, estimation and interval construction. In Section 6.3, we employ the proposed approach to model jointly glycated haemoglobin and fructosamine as flexible functions of covariates. Lastly, Appendix D includes the main code snippets used for the empirical analysis.

6.2 Copula Generalized Additive Models for Location, Scale and Shape (GGAMLSS)

Let us consider a pair of two continuous random variables, Y_1 and Y_2, and a generic covariate vector, $\boldsymbol{\nu}$. Note that we have suppressed observation index i for notational convenience, however recall that our aim is to model independent bivariate realizations $(y_{i1}, y_{i2})'$ as functions of covariates, where $i = 1, \ldots, n$ and n is the sample size. The joint cumulative distribution function (cdf) of Y_1 and Y_2 can be expressed in terms of the marginal cdfs of Y_1 , (i.e $F_1(y_1 \mid \mu_1, \sigma_1, \upsilon_1)$) and Y_2, (i.e. $F_2(y_2|\mu_2, \sigma_2, \upsilon_2)$) and a copula function $C(\cdot, \cdot, \rho)$ that binds them together - see Chapter 4. In this framework, $\mu_m, \sigma_m, \upsilon_m$, for $m = 1, 2$, are the marginal distribution parameters, and ρ is the association parameter measuring the dependence between the two random variables (Sklar, 1959).

In this chapter, we consider marginals with two and three parameters, but the formulation can be extended to parametric distributions with more parameters (see Chapter 4). Note that the parameters defined in $\vartheta = (\mu_1, \sigma_1, \upsilon_1, \mu_2, \sigma_2, \upsilon_2, \rho)'$, are linked to $\boldsymbol{\nu}$ via additive predictors as described in the next section.

The definitions of all marginal distributions implemented in `GJRM`-package are given in Table 2 of Marra and Radice (2017a) and are the normal (`"N"`), log-normal (`"LN"`), Gumbel (`"GU"`), reverse Gumbel (`"rGU"`), logistic (`"LO"`),

Weibull ("WEI"), inverse Gaussian ("iG"), gamma ("GA"), Dagum ("DAGUM"), Singh-Maddala ("SM"), beta ("BE"), and Fisk ("FISK"). The texts within the brackets are the values to use for the margins option in the gjrm() function. Table 1 of the same paper shows the copula functions implemented in the package, except for the Plackett which has been implemented for this work with the aim of extending the set of copulae available for practical modelling. The Plackett copula is $(Q-\sqrt{R})/\{2(\rho-1)\}$, where $Q = 1+(\rho-1)(u+v)$, $R = Q^2-4\rho(\rho-1)uv$ and $\rho \in (0,\infty)$. To sum up, the possible choices are the Gaussian ("N"), Clayton ("C0"), Joe ("J0"), Gumbel ("G0"), Frank ("F"), Ali-Mikhail-Haq ("AMH"), Fairlie-Gumbel-Morgenstern ("FGM"), Student-t ("T") and Plackett ("PL"). For Clayton, Gumbel and Joe, the number after the capital letter in the texts within the brackets indicates the degree of rotation required: the possible values are 0, 90, 180 and 270. The typewriter texts are the values to use for the BivD option in the gjrm() function. As shown in Chapter 4 and in Table 1 of Marra and Radice (2017a), there exist a relation between ρ and the well-known Kendall's $\tau \in [-1,1]$. This is useful since parameter ρ is often not easy to interpret, in which case the Kendall's τ can be used instead.

6.2.1 Flexible additive predictors

The framework adopted here allows one to relate all marginal distribution and dependence parameters to additive predictors (η's) via known monotonic link functions which ensure that the restrictions on the parameter spaces are maintained. As an example, if σ_1 can only take positive values and we wish to model this parameter as a function of covariates and regression coefficients then we can specify $g(\sigma_{1i}) = \eta_{\sigma_1 i}$, where the link function $g(\cdot)$ is equal to $\log(\cdot)$; see Marra and Radice (2017a) for details on the link functions. In this work, we assume a fully parametric specification for the distribution of the bivariate response vector (using parametric copulae and marginals as described in the previous section and Chapter 4) and that all the parameters of the bivariate distribution are related to regression coefficients and (e.g., binary, continuous and functional) covariates collected in $\boldsymbol{\nu}_i$ via an additive predictor generically (as defined in equation (4.3) of Chapter 4), i.e

$$\eta_i = \beta_0 + \sum_{j=1}^{J} f_j(\boldsymbol{\nu}_i), \tag{6.1}$$

where $\beta_0 \in \mathbb{R}$ is an overall intercept, $\boldsymbol{\nu}_i$ the covariate vector, and the J functions $f_j(\boldsymbol{\nu}_j)$ represent generic effects which are chosen according to the type of covariate considered. Note that to avoid clutter in the notation we have suppressed the subscript indicating which parameter the additive predictor belongs to. Each $f_j(\boldsymbol{\nu}_j)$ can be approximated as a linear combination of D_j basis functions $B_{j,d_j}(\boldsymbol{\nu}_i)$ and regression coefficients $\beta_{j,d_j} \in \mathbb{R}$ (as specified in equation (4.4) of Chapter 4). That is,

$$\sum_{d_j=1}^{D_j} \beta_{j,d_j} B_{j,d_j}(\boldsymbol{\nu}_i). \tag{6.2}$$

Equation (6.2) implies that the vector of evaluations $\{f_j(\boldsymbol{\nu}_1), \ldots, f_j(\boldsymbol{\nu}_n)\}'$ can be written as $\mathbf{Z}_j\boldsymbol{\beta}_j$ with $\boldsymbol{\beta}_j = (\beta_{j,1}, \ldots, \beta_{j,D_j})'$ and design matrix $Z_j[i, d_j] = B_{j,d_j}(\boldsymbol{\nu}_i)$. This means that equation (6.1) can be defined as

$$\boldsymbol{\eta} = \beta_0\mathbf{1}_n + \mathbf{Z}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{Z}_J\boldsymbol{\beta}_J, \tag{6.3}$$

as showed in equation (4.5) of Chapter 4. Equation (6.3) can also be written in a more compact way as $\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\beta}$, where $\mathbf{Z} = (\mathbf{1}_n, \mathbf{Z}_1, \ldots, \mathbf{Z}_J)$ and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_J')'$.

Each $\boldsymbol{\beta}_j$ has an associated quadratic penalty $\lambda_j\boldsymbol{\beta}_j'\mathbf{K}_j\boldsymbol{\beta}_j$ whose role is to enforce specific properties on the j^{th} function, such as smoothness. The smoothing parameter $\lambda_j \in [0, \infty)$ controls the trade-off between fit and smoothness, and plays a crucial role in determining the shape of $\hat{f}_j(\boldsymbol{\nu}_i)$. The overall penalty can be defined as $\boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta}$, where $\mathbf{K} = \text{diag}(0, \lambda_1\mathbf{K}_1, \ldots, \lambda_J\mathbf{K}_J)$.

An important feature of the approach adopted here is that many types of effects can be modelled in a unified manner. To date, the use of non-linear, random and spatial effects as well as interactions has been explored in copula-based regression models. However, functional effects have not been considered in the context of CGAMLSS. The following paragraphs outline the effects employed in our biomedical study.

Binary or categorical variables

For binary variables, predictor equation can be simplified as $\mathbf{z}_{ij}'\boldsymbol{\beta}_j$ and the design matrix is obtained by stacking all covariate vectors $\mathbf{z}_{ij}$ into $\mathbf{Z}_j$. No penalty is usually assigned to linear effects, hence $\mathbf{K}_j = \mathbf{0}$.

Continuous covariates

To model non-linear effects of the continuous covariates, different penalized spline definitions can be employed such as penalized low rank thin plate splines (Wood, 2003) and P-splines as proposed by Eilers and Marx (1996). For each continuous variable, ν_i, $f_j(\nu_i) = \sum_{d_j=1}^{D_j} \beta_{j,d_j} B_{j,d_j}(\nu_i)$, where $B_{j,d_j}(\nu_i)$ are known spline basis functions. In this chapter, we will employ thin plate regression splines which are numerically stable and have convenient mathematical properties. To enforce smoothness, a conventional integrated square second derivative spline penalty is employed.

In practice, we use `s()` when specifying the equations of `gjrm()` which essentially exploit the smoother set up of the `mgcv R` package. For instance, `s(z, bs = "tp", k = 10)`, where `z` is a continuous covariate, `k` is the number of

basis functions and argument `bs` specifies the type of spline basis (the default option is `t`hin `p`late regression `s`pline basis but other options are available).

Functional predictor

Assume, as in our case, that we have a functional covariate taking observed values $\{Z_i(t), t \in T\}$ for $i = 1, \ldots, n$, where Z_i is a square-integrable, random curve on the compact interval T. We assume that $Z_i(t) \in \mathbb{R}, \forall t, i$.

Here we can model the functional covariate into the CGAMLSS regression models as follows (McLean et al., 2014)

$$\int_T F\left[Z_i(t), t\right] dt, \tag{6.4}$$

where F is an unknown smooth surface to be estimated. Such effect can be estimated using integrating products of B-splines with roughness penalties over the functional covariates. Note that this new proposal using P-splines obey rather strong assumptions considering the linearity of the influence of the functional covariates.

More specifically, McLean et al. (2014) proposes to used a bivariate spline model to estimate F,

$$F(z,t) = \sum_{j=1}^{K_z}\sum_{k=1}^{K_t} \gamma_{j,k} B_j^Z(z) B_k^T(t), \tag{6.5}$$

where $B_j^Z(z)$ for $j = 1, \ldots, K_z$ and $B_k^T(t)$ for $k = 1, \ldots, K_t$ are spline basis on $[0, 1]$. If follows that equation (6.5) can be rewritten as

$$\int_T F\left[Z_i(t), t\right] dt = \sum_{j=1}^{K_z}\sum_{k=1}^{K_t} \gamma_{j,k} W_{j,k}(i), \tag{6.6}$$

where $W_{j,k}(i) = \int_T B_j^Z\left\{Z_i(t)\right\} B_k^T(t) dt$. In this expression, $W_{j,k}(i)$ can be approximated e.g. using the Simpson's rule. Furthermore, for identifiability, it is assumed $\sum_{i=1}^N \int_T F(Z_i(t), t) dt = 0$.

In addition, according to McLean et al. (2014) we have transformed the functional covariate $Z(t)$ using $G_t(z) = P\left\{Z(t) < z\right\}$ for each value of t. The use of this transformation ensures that each tensor-product B-spline has observed data on its support. Equation (6.6) can be expressed as

$$\int_T F\left[G_t\left\{Z_i(t)\right\}, t\right] dt = \sum_{j=1}^{K_z}\sum_{k=1}^{K_t} \gamma_{j,k} \int_T B_j^G\left[G_t\left\{Z_i(t)\right\}\right] B_k^T(t) dt, \tag{6.7}$$

where B_j^G is a new B-spline basis where for any t, the transformed points will be uniformly between $[0, 1]$. Function $G_t(z)$ can be estimated by using the empirical cumulative distribution function. The rest of the estimation procedure is analogous when this transformation is not used.

To enforce smoothness, penalties are considered. Let $\boldsymbol{W}_i$ the K_zK_t-dimensional vector formed by stacking the columns of $\mathbb{W}(i) = [W_{j,k}(i)]_{j=1,\ldots,K_z}^{k=1,\ldots,K_t}$, and let the matrix $\mathbb{W} = [\boldsymbol{W}_1\boldsymbol{W}_2...\boldsymbol{W}_n]'$.

Following Eilers and Marx (1996), the row penalty can be defined as $\lambda_1 \sum_{j=d+1}^{K_z} \left(\Delta_j^d \gamma_{j,k}\right)^2$. In this expression, $\Delta_j^d \gamma_{j,k}$ denotes the d-th difference by rows of the coefficients $\gamma_{j-d,k}, \ldots, \gamma_{j,k}$ (fixed k). Analogously, the column penalty is defined as $\lambda_2 \sum_{k=d+1}^{K_t} \left(\Delta_k^d \gamma_{j,k}\right)^2$ where $\Delta_k^d \gamma_{j,k}$ is the d-th difference of $\gamma_{j,k-d}, \ldots, \gamma_{j,k}$ where j is held fixed.

It follows that the penalty matrix can be expressed as $\mathbb{P} = \lambda_1 \mathbb{P}_1' \mathbb{P}_1 + \lambda_2 \mathbb{P}_2' \mathbb{P}_2$ where $\mathbb{P}_1 = \mathbb{K}_z \otimes \mathbb{I}_{K_t}, \mathbb{P}_2 = \mathbb{I}_{K_z} \otimes \mathbb{K}_t$; $\mathbb{I}_b$ is the $(b \times b)$-identity matrix, $\otimes$ denotes the Kronecker product, and $\mathbb{K}_z$ and $\mathbb{K}_t$ denote the matrix of row and column difference penalties, respectively. To take into account the intercept, a leading column of ones is added to matrix $\mathbb{W}$ and a first column of zeros is added to the penalty matrix $\mathbb{P}_1$ and $\mathbb{P}_2$.

Once the $\boldsymbol{\gamma}$ coefficient have been estimated, the estimated surface can be evaluated at any grid of points in its domain. Let $\boldsymbol{Z}$ a n_1-dimensional column vector of the functional covariate, and $\boldsymbol{T}$ the observation times n_2-dimensional vector taking values in $[0, 1]$. The estimated surface, $\widehat{\boldsymbol{F}}$, defined on $\boldsymbol{Z}$ and $\boldsymbol{T}$ is given by $\widehat{\boldsymbol{F}} = \mathbb{B}\widehat{\boldsymbol{\gamma}}_{[-1]}$, where $\widehat{\boldsymbol{\gamma}}_{[-1]}$ denotes the vector $\widehat{\boldsymbol{\gamma}}$ excluding first entry, which corresponds to the intercept, and $\mathbb{B}$ denotes the $n_1 n_2 \times K_z K_t$ matrix defined as $\mathbb{B} = \left(\mathbb{B}_z \otimes \mathbf{1}_{K_t}'\right) \odot \left(\mathbf{1}_{K_z}' \otimes \mathbb{B}_t\right)$. In this expression $\odot$ is the element-wise matrix multiplication and $\mathbf{1}_p$ denotes a column vector of length p. $\mathbb{B}_z$ is the $n_1 n_2 \times K_z$ matrix of z-axis B-splines evaluated at $\boldsymbol{Z} \otimes \mathbf{1}_{n_2}$. Analogous, $\mathbb{B}_t$ is the $n_1 n_2 \times K_t$ matrix of B-splines evaluated at $\mathbf{1}_{n_1} \otimes \boldsymbol{T}$.

To illustrate this procedure into CGAMLSS framework, let `Glu` be the 560×36 matrix, with one glucose profile per row, and `Time` be the matrix of times at which the glucose values are measured (all rows are identical). That is,

```
> head(Time[, 1:5])
0.08333333 0.1666667 0.25 0.3333333
0.08333333 0.1666667 0.25 0.3333333
0.08333333 0.1666667 0.25 0.3333333
0.08333333 0.1666667 0.25 0.3333333
0.08333333 0.1666667 0.25 0.3333333
```

```
> head(Glu[, 1:5])
Glu(t11)  G2(t12)   G3(t13)  G4(t14)
127       128       130      134
131       131       130      129
138       136       133      131
107       107       109      110
87        88        90       93
```

```
> n=nrow(Glu)
> nt=ncol(Glu)
> L=((Time[, nt] - Time[, 1])/nt)/3 * matrix(c(1, rep(c(4,
2), length = nt - 2), 1), nrow = n, ncol = nt, byrow = T)
```

Note that `Glu` and `Time` have identical dimensions and `L` denotes the 560 $\times$ 36 matrix of quadrature weights to use in the numerical integration of the surface F. The smooth formula one has to use to adjust for such functional effect is `s(z = Time, x = Glu, by = L, bs = "dt", xt = list(tf = list(Glu = "QTransform"), basistype = "te"))`. Functional terms can be included in all the equations of the bivariate copula model. In addition, several functional covariates can be considered as predictors. Furthermore, these terms can be easily mixed with other types of smooths (used to model non-linear, spatial and random effects, for instance).

6.2.2 Estimation and inferential details

The log-likelihood function for a copula model with continuous margins can be written as

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} log\left\{c(F_{1i}(y_{1i}|\mu_{1i},\sigma_{1i},\upsilon_{1i}), F_2(y_{2i}|\mu_{2i},\sigma_{2i},\upsilon_{2i});\rho_i)\right\} + \sum_{i=1}^{n}\sum_{m=1}^{2} log\left\{f_m(y_{mi} \mid \mu_{mi},\sigma_{mi},\upsilon_{mi}))\right\}. \tag{6.8}$$

In the above equation, the distributional parameters are defined as follows: $\mu_{mi} = g_{\mu_m}^{-1}(\eta_{\mu_m i})$, $\sigma_{mi} = g_{\sigma_m}^{-1}(\eta_{\sigma_m i})$, $\upsilon_{mi} = g_{\upsilon_m}^{-1}(\eta_{\upsilon_m i})$, for $m = 1, 2$, $\rho_i = g_{\rho}^{-1}(\eta_{\rho i})$, the g's are link functions, $c(\cdot,\cdot,\rho)$ is the density function of the copula function, and $f_m(y_m|\mu_m,\sigma_m,\upsilon_m)$ the density of the m^{th} marginal. Parameter vector $\boldsymbol{\theta}$ is defined as $(\boldsymbol{\beta}'_{\mu_1},\boldsymbol{\beta}'_{\mu_2},\boldsymbol{\beta}'_{\sigma_1},\boldsymbol{\beta}'_{\sigma_2},\boldsymbol{\beta}'_{\upsilon_1},\boldsymbol{\beta}'_{\upsilon_2},\boldsymbol{\beta}'_{\rho})'$ which refer to the coefficient vectors associated with $\eta_{\mu_1 i}$, $\eta_{\mu_2 i}$, $\eta_{\sigma_1 i}$, $\eta_{\sigma_2 i}$, $\eta_{\upsilon_1 i}$, $\eta_{\upsilon_2 i}$, and $\eta_{\rho i}$.

Because of the presence of flexible additive predictors in the model, estimation is carried out within a penalized likelihood-based framework. Specifically, we maximize

$$\ell_p(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}'\mathbf{S}\boldsymbol{\theta}, \tag{6.9}$$

where ℓ_p is the penalized log-likelihood of the model, and $\mathbf{S} = \text{diag}\,(\mathbf{K}_{\mu_1},\mathbf{K}_{\mu_2},\mathbf{K}_{\sigma_1},\mathbf{K}_{\sigma_2},\mathbf{K}_{\upsilon_1},\mathbf{K}_{\upsilon_2},\mathbf{K}_{\rho})$. The overall smoothing parameter is defined as $\boldsymbol{\lambda} = (\boldsymbol{\lambda}'_{\mu_1},\boldsymbol{\lambda}'_{\mu_2},\boldsymbol{\lambda}'_{\sigma_1},\boldsymbol{\lambda}'_{\sigma_2},\boldsymbol{\lambda}'_{\upsilon_1},\boldsymbol{\lambda}'_{\upsilon_2},\boldsymbol{\lambda}'_{\rho})'$. In practice, estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ is achieved by using the stable and efficient trust region algorithm with integrated automatic multiple smoothing parameter selection by Marra and Radice (2017a).

At convergence, reliable point-wise confidence intervals for linear and non-linear functions of the model coefficients can be obtained using the Bayesian

large sample approximation (Marra and Radice, 2017a)

$$\boldsymbol{\theta} \dot{\sim} N(\hat{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}}), \tag{6.10}$$

where $\hat{\boldsymbol{\theta}}$ is a parameter vector estimate, $\mathbf{V}_{\boldsymbol{\theta}} = -\boldsymbol{H}_p(\hat{\boldsymbol{\theta}})^{-1}$ and $\boldsymbol{H}_p$ is the penalized model's Hessian. Intervals derived using (6.10) have good frequentist properties since they account for both sampling variability and smoothing bias (e.g., Marra and Radice, 2017a, and references therein). Intervals for non-linear functions of the model's coefficients (e.g., τ, joint and conditional predicted probabilities) can be conveniently obtained by simulation from the posterior distribution of $\boldsymbol{\theta}$ using the following steps:

- Draw n_{sim} random vectors from $N(\hat{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}})$.
- Calculate n_{sim} simulated realizations of the quantity of interest. For instance, for a Gaussian copula we have that $\tau_i = \frac{2}{\pi} \arcsin \{\tanh (\eta_{\rho i})\}$. Vector $\boldsymbol{\tau}_i^{sim} = (\tau_i^{sim_1}, \tau_i^{sim_2}, \ldots, \tau_i^{sim_{n_{sim}}})' \; \forall i = 1, \ldots, n$ is obtained using $\boldsymbol{\beta}_{\rho}^{sim_j}$ $\forall j = 1, \ldots, n_{sim}$ and the transformation just described.
- For each $\boldsymbol{\tau}_i^{sim}$, calculate the lower, $\varsigma/2$, and upper, $1 - \varsigma/2$, quantiles.

A small value for n_{sim}, say 100, typically gives accurate results (although it can be increased if more precision is desired), whereas ς is usually set to 0.05.

The discussed copula models can be easily fitted in `R` using the package `GJRM` (Marra and Radice, 2017b). For instance,

```
fl <- list(y1 ~ x1 + s(x2) + s(x2, by = x3), y2 ~ x1
+ s(x2))
md <- gjrm(fl, margins = c("LO", "WEI"), BivD = "PL",
Model = "B")
```

where `fl` is a list containing (in this simple case) two equations, `margins` specifies the marginal distributions and `BivD` the copula. Argument `Model = "B"` means that a bivariate model will be employed (other models are available).

6.3 Modelling jointly HbA1c and fructosamine

The aim of this investigation was to model jointly HbA1c and fructosamine as function of flexible covariate effects in a population-based study taking into account the glucose profiles of the individuals. In the following sections we describe the data, the model building process and comment on the results obtained.

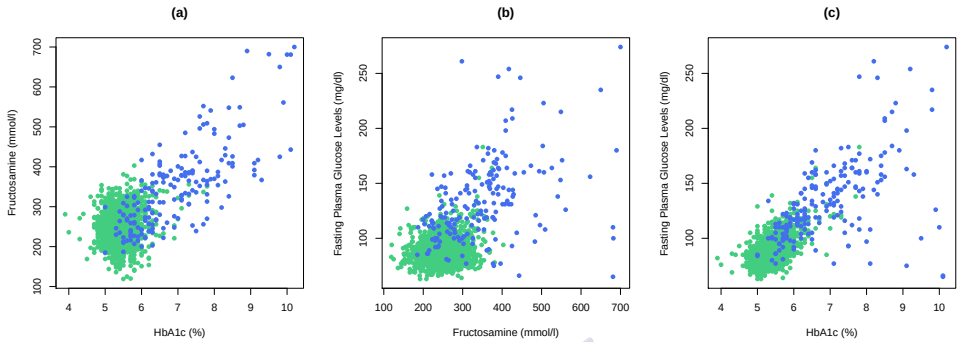


Figure 6.1: a) Scatterplot between glycated haemoglobin (HbA1c) and fructosamine. b) Scatterplot of fructosamine against glucose. c) Scatterplot of HbA1c against glucose. In all these plots, people with and without diabetes are represented with different colours.

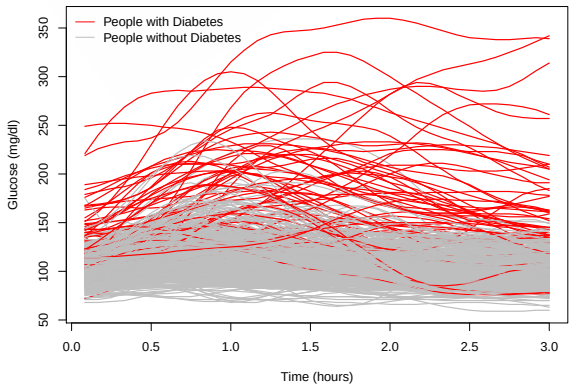


Figure 6.2: Glucose profiles from 560 individuals collected on the third day, over three hours after the breakfast.

6.3.1 Data description

Data come from the AEGIS study (sub-project CGM). See Chapter 4 for a description of the study.

To obtain a more accurate measurement of the glucose levels over time, doctors frequently performed blood tests to assess blood glucose levels. There were two common lab tests performed to check blood glucose: HbA1c and fructosamine. Both are based on a relevant phenomenon called "glycation," a process by which glucose is chemically bound to amino groups of proteins. However, the exact contribution of postprandial blood glucose excursions to the overall glycemic control of patients with type 2 diabetes remains largely undetermined. Thus, we were interested in the effect of postprandial glucose excursions on the levels and variability of both glycated proteins, knowing that both responses are highly correlated depending on the levels of fasting plasma glucose (see Figure 6.1).

In this study, we used the values recorded every 5 minutes along three hours after breakfast on the third day of monitoring. The final number of samples used in the analysis was 560. The following covariates were considered: age (in years), continuous interstitial glucose monitoring displayed in Figure 6.2 and called Glu in the next section, and fasting plasma glucose of the individuals (fpg). The HbA1c and fructosamine levels made up the bivariate response.

6.3.2 Model building

In this section, we describe the process used for building the bivariate model for the joint distribution of HbA1c and fructosamine. The response distributions and copula were chosen using the Akaike and Bayesian Information Criteria (AIC and BIC) together with normalized quantile residuals (Marra and Radice, 2017a). To simplify the model building process, we exploited the fact that in a copula context the specification of margins and copula can be viewed as separate but related issues.

A parsimonious model was specified using both clinical knowledge and results available in the literature. These show that age and plasma glucose are important factors affecting the mean HbA1c and fructosamine concentrations (Nakashima et al., 1993; Pani et al., 2008; Monnier and Colette, 2009). In addition, it is here hypothesized (since the literature contains no information in this respect) that age influences the variance of both glycated proteins. Other variables may also have an influence (although to a smaller extent) on glycation levels, including the BMI and aspects of kidney function that are not taken into account here for the sake of simplicity.

On the basis of the data available as well as our main research question (i.e., investigate the effect of postprandial glucose on the levels of HbA1c and fructosamine), the additive predictors for the location parameters of the marginal

distributions were specified as

$$\eta_i = \beta_0 + \int F_1(\text{Glu}_i(t), t)dt + f_2(\text{age}_i),$$

whereas for the scale parameters the additive predictors were defined as

$$\eta_i = \beta_0 + f_1(\text{age}_i).$$

Finally, knowing that both diabetic proteins are highly correlated and dependent on the levels of fpg, as mentioned in the previous section, the additive predictor for the copula parameter was defined as

$$\eta_i^\rho = \beta_0^\rho + f_1^\rho(\text{fpg}_i).$$

The smooth functions of age and fpg capture the possibly non-linear effects of these variables, and F_1 is the functional effect of the glucose variable.

We first chose the marginal distributions. The CGAMLSS approach can handle several distributions for the responses, the majority of which were contemplated in this study. We did not consider three parameter distributions as these typically require larger sample sizes to produce reliable results. Using the AIC, BIC as well as Q-Q plots of the normalized quantile residuals, we arrived at the logistic and log-normal distributions for HbA1c and fructosamine, respectively. Table 6.1 shows the AIC and BIC values for the candidate marginal distributions, whereas Figure 6.3 shows Q-Q plots of normalized quantile residuals for the chosen distributions. The plots show that, overall, the chosen distributions fit the data well. However, there are some departures from the reference line for higher values of these variables.

As for the choice of copula, we started off with the Gaussian and then, based on the (negative or positive) sign of the dependence, we tried out alternative specifications. In this case, the values for the correlation and τ coefficients were found to be both positive and negative. Therefore, we only considered copula which were consistent with this finding. For example, the Gaussian, Frank, AMH and FGM copulae allow for both positive and negative dependence. Frank exhibits a slightly stronger dependence in the middle of the distribution as compared to the Gaussian. There also exist asymmetric copulae as Clayton, Joe or Gumbel. Clayton is asymmetric with a strong lower tail dependence but a weaker upper tail dependence. The opposite is true for the Gumbel and Joe. AMH, FGM and Plackett can only account for weak dependencies. In this study, the use of Clayton, Joe, Gumbel and Plackett led to convergence failure, suggesting that these copulae are not appropriate for this case study.

According to expert physician knowledge, and the statistical criteria mentioned above (AIC and BIC), the best fit was provided by the Gaussian copula (see Table 6.2). The results obtained with this copula support the hypothesis that the highest levels of haemoglobin glycation are associated with high postprandial plasma glucose. The proposed model allows the glycation levels of both proteins to be studied.

Table 6.1: Comparison of AIC and BIC values for the candidate marginal distributions for HbA1c and fructosamine.

	Y_1: HbA1c		Y_2: fructosamine	
Distributions	AIC	BIC	AIC	BIC
Normal	141.20	209.57	4998.38	5104.58
Gumbel	251.22	312.05	5056.08	5176.94
Reverse Gumbel	194.74	247.15	5048.46	5152.65
Logistic	101.73	142.83	5002.33	5108.37
Log-normal	142.03	210.99	4989.56	5095.65
Weibull	219.27	278.25	5008.82	5130.51
Inverse Gumbel	490.93	5004.15	5112.63	5931.52
Gamma	277.76	308.32	5845.23	5932.21

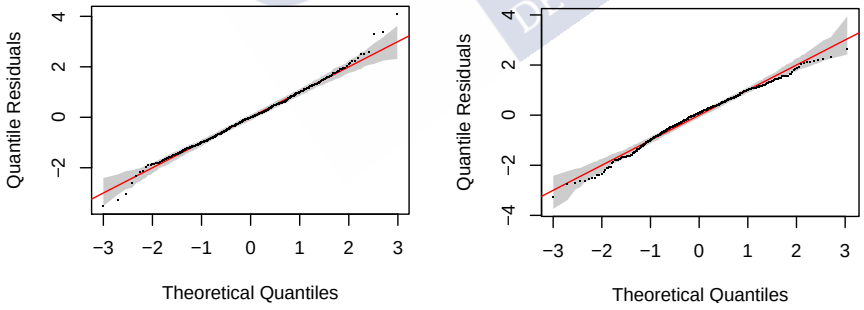


Figure 6.3: Q-Q plots of normalized quantile residuals for HbA1c (left) and fructosamine (right) produced after fitting a Gaussian copula model with logistic and log-normal margins to the AEGIS data. Note that the Q-Q plots also exhibit reference bands for judging the relevance of departures from the red reference lines. In these cases, the distributions fit the main bulk of the data well. However, there are some departures for higher values of these variables.

Table 6.2: Comparison of AIC/BIC choice criteria under some copula assumptions.

Copula	AIC	BIC
Gaussian	5017.39	5190.93
Gumbel	5105.66	5265.33
Frank	5100.77	5300.00
AMH	5127.39	5436.45

Table 6.3: Definition and some of the properties of the distributions used in the case study. $\text{erf}(\cdot)$ denotes the error function. Note that for both distributions μ can take any value on the real line whereas σ can only take positive values.

$f(y\|\mu,\sigma)$		$F(y\|\mu,\sigma)$	$\mathbb{E}(y)$	$\text{Var}(y)$
Log-Normal	$\frac{1}{y\sigma\sqrt{2\pi}}\exp\left[-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right]$	$\frac{1}{2}+\frac{1}{2}\text{erf}\left\{\frac{\log(y)-\mu}{\sigma\sqrt{2}}\right\}$	$\sqrt{\exp(\sigma^2)}\exp(\mu)$	$\exp(\sigma^2)\{\exp(\sigma^2)-1\}\times$ $\exp(2\mu)$
Logistic	$\frac{1}{\sigma}\left\{\exp\left(-\frac{y-\mu}{\sigma}\right)\right\}\times$ $\left\{1+\exp\left(-\frac{y-\mu}{\sigma}\right)\right\}^{-2}$	$\frac{1}{1+\exp\left(-\frac{y-\mu}{\sigma}\right)}$	μ	$\frac{\pi^2\sigma^2}{3}$

6.3.3 Empirical results

Given the complexity of the model employed in this study, special attention needs to be devoted to the clinical interpretation of results. In a classical regression model, based for instance of the assumption of normality, it is sufficient to discuss the estimated effects directly. However, when using non-Gaussian distributions, quantities like expectation and variance may be determined through functions of the distribution's parameters. Table 6.3 shows the properties of the distributions employed in our biomedical study; these have been used to calculate the covariate effects. As for the functional effects, here we display some plots which can aid their interpretation. The newly introduced functions `pred.mvt` (which allows one to predict the mean and variance of a given marginal distribution) and `vis.gjrm` (which produces perspective or contour plots views of the model predictions) are available in `GJRM` and can be used to make more sense of the results. See Appendix D for more details.

Figures 6.4, 6.5 and 6.6 display the effects of the covariates on the mean and variance of the responses, and on the association between them. For Figures 6.4(a), 6.4(b), 6.4(d), 6.4(e), 6.5, variable age was set to 46 years (the mean), respectively. For Figures 6.4(c), 6.4(f), variables age, *glu* and *time* were set to 46 years, 108 mg/dL and 0.5 hours, respectively. Finally, for Figure 6.6, variables age, Glu, time were set to 46 years, 108 mg/dL, 0.5 hours. The findings can be summarised in three main blocks

- Figure 6.4, marginal expectations. Mean HbA1c concentrations increase almost linearly with age. Instead, fructosamine concentrations only do so for elderly people (> 60 years). For the fructosamine, the confidence intervals

are wide which also suggest that age may not play an important role in explaining the response.

The age-related increase in HbA1c that we observe in our study is similar in magnitude to what found in a previous study analyzing data from 2473 nondiabetic participants of the Framingham Offspring Study (FOS), and in 3270 non-diabetic participants from the National Health and Nutrition Examination Survey (NHANES, 2001-2004) (Pani et al., 2008). As age increases the levels of HbA1c become larger in individuals, after adjusting for glucose levels. This suggests that factors unrelated to glucose's metabolism are affecting HbA1c levels. One such factor may be the change in the rate of glycation associated with aging (Nakashima et al., 1993). This article has also shown clear differences when comparing HbA1c with fructosamine in subjects of varying ages. Some of this inconsistency can be explained by the fact that HbA1c reflects glycaemic control over the preceding 6-8 weeks compared with 1-3 weeks for fructosamine. Fructosamine measurements can also be influenced by other factors such as the serum albumin concentration and body mass index of the subject.

As for the functional covariate, the levels of both proteins are higher for higher levels of glucose at one hour post-meal. However, the time dependent effect of glucose on glycated protein levels seems to be more pronounced for HbA1c than fructosamine. This interesting finding does not agree with the hypothesis that HbA1c is an indicator which mainly reflects mean plasma glucose level but does not reflect postprandial plasma glucose well; and that glycated albumin (the main component of fructosamine) correlates most closely with the postprandial glucose levels (Sakuma et al., 2011). HbA1c, which remains the gold standard for assessing glucose homeostasis, is an integration of both fasting and postprandial glucose variations over a 3-month period, in which the respective contributions of both fasting and postprandial glucose are still a subject of debate (Monnier and Colette, 2009).

In recent years, new data have provided further information on the ongoing debate as to whether HbA1c, fasting glucose and postprandial glucose contribute equally or not to the overall hyperglycemia in type 2 diabetes. While some studies reported that preprandial plasma glucose concentrations are related to HbA1c more strongly than with postprandial concentrations (Bonora et al., 2001). Avignon et al. (1997) found that post-lunch plasma glucose values correlate better with overall glycemic control as estimated from HbA1c than do pre-breakfast and pre-lunch glucose levels. In an analysis of a dataset collected in the Diabetes Control and Complications Trial, Rohlfing et al. (2002) reported that a better association with HbA1c was obtained for post-lunch and mean daily glucose concentrations. Our results support the hypothesis that the highest glycation rates

of hemoglobin are a consequence of high glucose levels, which appear in the postprandial period.

- Figure 6.5, marginal variances. The variance plots suggest that variation in HBA1c and fructosamine seems to be not related to age. Unfortunately, information regarding whether the variation of glycation proteins is different along age is not available yet.

- Figure 6.6, dependence. This Figure shows that the correlation between glycated hemoglobin and fructosamine is lower in subjects with fasting plasma glucose levels aroung $90 - 120$ mg/dL than in subjects at lower (fpg ≈ 80 mg/dL) and higher (fpg > 130 mg/dL). Glucose attaches non-enzymatically to amino groups of proteins to form glycated haemoglobin or ketoamines. Thus it is expected to find higher glycation rates at higher glucose levels and therefore higher correlation between both glycated proteins at higher glucose levels. The greater correlation between both glycated proteins at lower glucose levels may be explained because hypoglycaemia is an adverse event in patients with diabetes who receive insulin treatment. From a clinical point of view, these results suggest in some patients that the correlation between both proteins is greater in patients with diabetes than in subjects without diabetes.

The plots in this section have been produced using the newly introduced functions `pred.mvt` and `vis.gjrm` in `GJRM` which make use of the results in Table 6.3; see the Appendix D for more details.

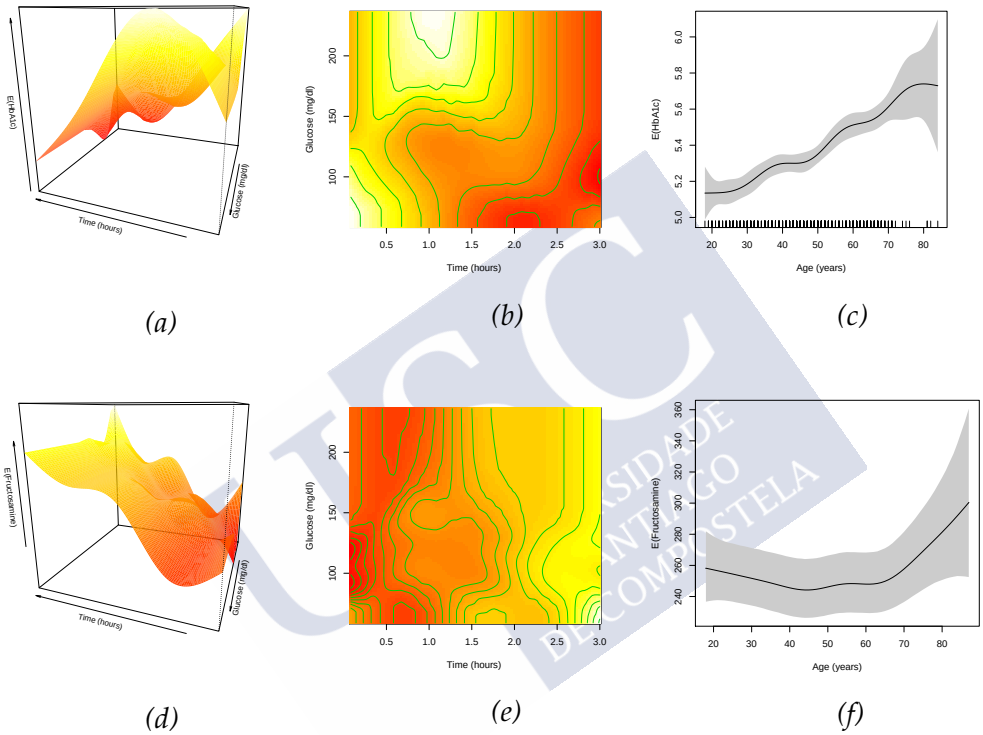


Figure 6.4: Estimated smooth effects of glucose over the time, and age on the mean of HbA1c (top plots) and fructosamine (bottom plots) obtained when fitting a Gaussian copula model with logistic and log-normal margins for HbA1c and fructosamine, respectively. Figures (a) and (d) show perspective plots of the glucose effect over time, whereas Figures (b) and (e) display the contour plots for the same effect. In the contour and perspective plots, red corresponds to small mean levels and yellow to high ones.

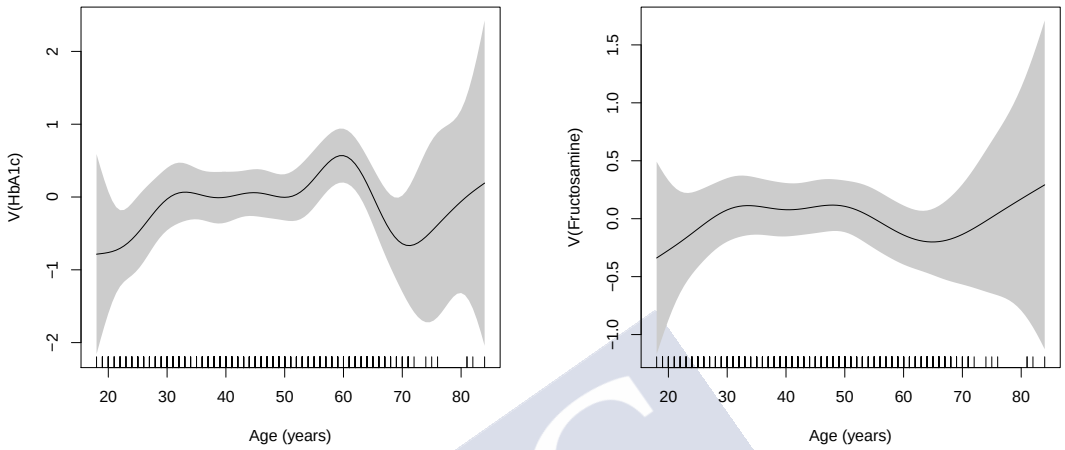


Figure 6.5: Estimated smooth effects of age on the variance of HbA1c (right plot) and fructosamine (left plot) obtained when fitting a Gaussian copula model with logistic and log-normal margins for HbA1c and fructosamine, respectively.

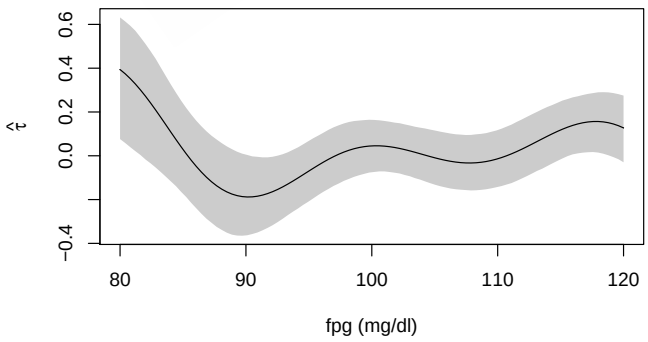


Figure 6.6: Estimates and 95% intervals for τ from a Gaussian copula model with logistic and log-normal margins for HbA1c and fructosamine, respectively.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 7

Discussion and future research

In this chapter, we present a summary of the main results obtained in this thesis (Chapters 3-6) together with some conclusions and interesting lines for future research.

7.1 Chapter 3: "Detecting differences in blood potassium concentrations by using a spatial distributional regression model"

Distributional regression models extends the use of Generalized Additive Models (GAM, Hastie and Tibshirani, 1990) to situations in which the response distributions are non-standard, and in which not only the mean but multiple parameters are related to additive predictors via suitable link functions. Further, they allow additional flexibility by specifying structured additive predictors for each parameter of interest, and thus adjust for flexible non-linear effects of continuous covariates for which the smoothness is determined based on the data. They also allow the contemplation of spatial effects to capture unobserved spatial heterogeneity and spatial correlations, interaction terms such as varying coefficients or interaction surfaces, and cluster-specific random effects (Fahrmeir et al., 2013; Brezger and Lang, 2006).

The use of a new structured additive distributional regression model allowed for the flexible modelling of the distribution of potassium concentrations with a potentially non-standard response type, and permitted covariate effects to be taken into account, including the smooth estimation of the effect of continuous variables, categorical covariates, random effects and possible spatial trends. Its use clearly identified differences in serum potassium concentrations among extraction sites after adjusting for other potentially influential factors, such as age, gender and clot-contact time. The spatial analysis revealed some districts to return higher mean serum potassium concentrations, and to show greater variability in terms of these results. Two geographically-related clusters were detected:

i) districts on the periphery of the study area that returned higher potassium concentrations (and showed greater variability in the results) than those in the central area, and ii) a number of districts that returned higher potassium concentrations independent of their location.

Although classic regression analyses allow for the easy interpretation of results, they only focus on means, and may lead to erroneous conclusions when modelling complex data structures. The distributional regression models used in this chapter provide a generic framework for performing regression analyses in which several parameters of a potentially non-standard response distribution are related to flexible regression predictors. Chapter 3 shows how to visualize from a statistical viewpoint the results of distributional regression models in an analysis comprising spatial information. It is not sufficient to show the estimated spatial effects directly; rather the spatial effect has to be adjusted with respect to the covariates observed in the particular regions. We therefore plugged in covariate values obtained as spatially stratified averages for all other covariates and then compared against the overall mean of the spatial effect to determine significances.

Another benefit of this type of model is the possibility of being able to contemplate a wide family of response variables. One way of examining the goodness of fit of the selected model is via quantile-quantile plots. However, conclusions drawn from such plots can be subjective. This thesis proposes the use of quantile-quantile plots with reference bands.

The present work examined different (non-standard) distributions that depended on the mean and variance of responses. Using the DIC, distributional regression with a log-normal response was used. This provided not only improved goodness of fit over classic distributions (e.g., a Gaussian distribution), but led to different results being obtained. Thus, although the non-linear effects of the covariates on the expectation were rather similar with both distributions, the log-normal distribution allowed differences to be identified in the variability of the spatial effects associated with the potassium concentrations that were undetectable when a normal distribution was contemplated (data not shown). The majority of the central municipalities had larger populations, more health care personnel and more equipment, and followed protocols more strictly, which might explain the lower potassium concentrations they recorded (with both distributions) and their smaller variability (log-normal distribution). In medical organizations, examining clinical variation in medical practice is an important step to measuring efficiency and effectiveness in care delivery.

From a statistical viewpoint, another important feature of this type of model, is the possibility of modelling the effects of the continuous covariates and spatial effects in a flexible, unified manner (as shown in the present work) as well as allowing for complex interactions between different types of variables, e.g., factor-curve or surface interactions. They also allow for the modelling of spatio-temporal trends. One of the hypotheses of the present work was that the holiday

periods of the extraction personnel might have an effect on a number of preanalytical factors (e.g., a greater chance of heamolysis occurring when less experienced personnel perform extractions). The present work does not contemplate such effects since data were available only for a short period (6 months). Future work will include extending the observation period to 5 years, allowing these spatio-temporal effects to be taken into account.

Further work is also required to determine whether the elevated potassium concentrations detected reflect a real clinical panorama or a problem of pseudohyperkalaemia. The latter scenario would appear to be more likely, however, since neither the lifestyles of those living in high value districts, nor the prevalence of disease in these areas, would seem able (at least on first inspection) to explain them. If the high values do reflect a pseudohyperkalaemia problem, a number of actions could be undertaken to help rectifying it, including: i) the education of laboratory and non-laboratory personnel about the causes of increased potassium readings; ii) the teaching of procedures to reduce the problem; iii) improving the transport routes to the hospital to reduce clot-contact times; and iv) constant monitoring of potassium concentrations and the apparent rate of hyperkalaemia. It is worth noting that the efficient management of laboratories and other health care services has received considerable attention in the optimization literature (e.g. Green, 2006; Mankowska et al., 2014). An investigation into the optimization techniques most appropriate for the present context might help curb the possible inefficiencies in the sample routing system.

Given the close relationship between potassium and sodium ions, it would be of great interest to study both cations at the same time in order to determine the covariates that influence them and their interactions. For this, flexible copula distributional regression models for multivariate responses can be used both in a Bayesian framework (structured additive conditional copula regression models Klein and Kneib, 2016b) or in a frequentist framework, using bivariate copula additive models for location, scale and shape (Marra and Radice, 2017a).

Finally, we would like to point out that this chapter contains some statistical improvements in the context of distributional regression. Specifically, two methodological innovations were included:

- The first methodological innovation concerns the construction of reference bands that were added to the quantile-quantile plot for determining the fit of a DR model. These proposed reference bands are obtained following Augustin et al. (2012) after the appropriate adjustment to the distributional regression context, see Section 3.3 in the chapter.

- Second, here is illustrated how to visualize the results of distributional regression models in an analysis comprising spatial information. In this case, it is not sufficient to show the estimated spatial effects directly but one rather has to adjust the spatial effect with respect to the covariates observed in the particular regions. We therefore plugged in covariate values obtained

as spatially stratified averages for all other covariates and then compared against the overall mean of the spatial effect to determine significances. In this framework, the designed code is provided to the reader; (see Appendix A).

7.2 Chapter 4: "Extensions to bivariate responses: Copula regression models"

Chapter 4 reviews and compares flexible modern strategies for simultaneously investigating factors that influence the discordance between markers used to screen for, and diagnose, individuals with diabetes. The performance of different bivariate CGAMLSS, based on either frequentist (Marra and Radice, 2017a) or Bayesian (Klein and Kneib, 2016b) inferential principles, was examined via a simulation and a real biomedical study. To the best of our knowledge this is the first time that these methodologies have been compared. The chapter shows that the compared methodologies offered similar results.

CGAMLSS is a new methodology that until now has been little employed in the biomedical setting. One of the reasons for this is that the results it provides are hard to interpret. The present chapter, however, highlights the value of CGAMLSS to medical researchers when dealing with datasets in which multivariate dependence is of interest and marginal distributions may come from different non-standard families. This work also shows how to visualize the results of CGAMLSS at the real scale of the response variables.

From a statistical standpoint, the CGAMLSS regression models reviewed provide a generic framework for performing regression analyses in which the parameters of a potentially non-standard multivariate response distribution are related to flexible regression predictors. An important feature of this type of model is the possibility of modelling, in a flexible and unified manner, different types of effect, such as spatio-temporal trends, interactions and random effects, as covariates. In addition, and as shown in the present thesis, they allow the non-linear effects of continuous covariates to be investigated. In the present work, penalized splines were employed, using the same number of knots, to model continuous covariates in both the frequentist and Bayesian approach (Eilers and Marx, 1996). However, other penalized spline definitions could be employed in the frequentist method, such as penalized low rank thin plate splines (Wood, 2003) or cubic regression splines (Wood, 2006). In the present study, the use of additive instead of linear predictors was particularly useful in detecting the effect of age and glucose concentration on the variability of the HbA1c and fructosamine values.

Therefore, this methodology is potentially very useful in biomedical research. In the case examined here, we have found that two biomarkers that are indistinctly used in diabetes control (HbA1c and fructosamine) can diverge in their

results depending on the characteristics of the patients. The most important clinical contribution of these models arises when studying the association (dependency) of these two response variables in light of the covariates. For example, thanks to the models, our results demonstrate that these two diagnostic tests show bigger discrepancies between them when the patients are young and their glucose levels are normal (Figures 4.14 and 4.15). This means that the interpretation of the results should take into account the individual characteristics of the patient under examination. In other words, any of the two biomarkers could be indistinctly used provided that the association between them is high and that there are no discrepancies when taking the covariates into account.

Another benefit of CGAMLSS is that it can contemplate a broad family of non-standard response variables. Although this study focuses on two continuous response variables, this framework allows one to estimate bivariate regression models with binary responses (where link functions are not restricted to probit alone) or bivariate models with binary/discrete/continuous margins in the presence of associated responses/endogeneity. In fact, the author of this thesis together with other investigators are currently investigating the concordance of the different diagnostic criteria for diabetes. Two of these criteria, established by American Diabetes Association (ADA), are fasting plasma glucose levels ($\geq$ 126 mg/dL) and $(HbA1c \geq 6.5\%)$ (American Diabetes Association, 2018). We are working on the development of a bivariate binary model that allows one to investigate whether concordance exists between these two diagnostic criteria, and whether the threshold levels used are the most appropriate. Other types of diagnostic criteria have also been defined by the ADA such as the 2-hours plasma glucose value during a 75-g oral glucose tolerance test. We are also working on the development of copula regression models for trivariate responses with a view to simultaneously studying the effectiveness of these three criteria, in the line with the recent proposal of Filippou et al. (2017). Preliminary results suggest that these approximations may help improve the diagnosis of diabetes.

Finally, we would like to point that these modern regression techniques may be useful for clinicians since they allow for simultaneously explain which mechanisms are affecting on multivariate responses and therefore using these models could shed lights on certain important biological processes. In this thesis, the usefulness of this methodology was tested in the setting of diabetes research, but it could be similarly used in studies on the markers of cancer, cardiovascular disease, and for instance in other studies of health related quality of life (Espasandín-Domínguez et al., 2018b). Further work is needed to bring this type of analysis into the clinic.

7.3 Chapter 5: "Distributional regression models including functional data"

In this chapter, we have presented the extension of DR regression models to consider functional data covariates. The flexibility of the approach allows to combine functional data covariates with another type as categorical, non-linear effects of continuous covariates or spatial effects, among others. Furthermore, there are a large number of non-Gaussian responses available in BayesX software that could be considered in this framework. Indeed, the current implementation of the software supports more than 12 continuous response distributions and 9 discrete distributions. In this setting, all parameters of these distributions can be related to additive predictors.

The present work, such as the most methods in functional data regression literature involved functional linear regression. Further work is need to develop DR extensions to consider nonlinear functional predictor covariates. Chapter 6 discusses this type of methods in more detail.

The proposal extension have been applied to study the determinant factors in differential fructosamine levels. We believe that the application of this methodology will be very useful in another biomedical studies. Furthermore, in future research, it would be interesting to simultaneous modelling more than one response distribution. See Chapter 6 for more details.

7.4 Chapter 6: "Functional regression CGAMLSS"

In this chapter, we have proposed flexible regression copula models with linear, non-linear and functional covariate effects to model jointly two diabetic proteins. The adopted approach allowed us to model flexibly the parameters of the non-Gaussian distributions chosen for HbA1c levels and fructosamine concentrations, the copula parameter describing the dependence between the responses, and to account for complex covariate effects. We have also provided software for the implementation of such models, created functions to aid the interpretation of results. To the best of our knowledge, this is the first study that has considered copula additive regression models with functional covariates of the same flexibility.

The CGAMLSS framework discussed in this chapter provides a generic framework for performing regression analyses in which several parameters of two potentially non-standard response distributions are related to flexible additive predictors. For instance, it would be also interesting to investigate the effect of postprandial glucose excursions on the levels and variability of both markers of glycemic control, after adjusting for the effect of con-founders such age as well as accounting for the way age and fasting plasma glucose modify the relationship between the two proteins. These findings may have important clinical implica-

tions when using glycated proteins as indicators of glycemic control in patients with diabetes.

As a possible improvement, we will look into extending the scope of the model specification considered in this chapter to accommodate all the measurements collected from the study participants over all days. Joint modelling of three continuous responses could also be of interest. Specifically, in this study we have considered two glucose markers, but given the close relationship between HbA1c, fructosamine and other proteins such as glycated albumin, it would be of great interest to study these three proteins simultaneously to assess which covariates influence the markers and how their dependence is modified by covariate effects.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

References

Acar, E. F., Craiu, R. V., and Yao, F. (2013). Statistical testing of covariate effects in conditional copula models. *Electronic Journal of Statistics*, 7:2822–2850.

American Diabetes Association (2001). Postprandial blood glucose. *Diabetes Care*, 24:775–778.

American Diabetes Association (2018). Classification and diagnosis of diabetes: Standards of medical care in diabetes - 2018. *Diabetes Care*, 41:S13–S27.

Antoniadis, A. and Sapatinas, T. (2003). Wavelet methods for continuous-time prediction using Hilbert-valued autoregressive processes. *Journal of Multivariate Analysis*, 87:133–158.

Augustin, N., Sauleaub, E., and Wood, S. (2012). On quantile-quantile plots for generalized linear models. *Computational Statistics and Data Analysis*, 56:2404–2409.

Avignon, A., Radauceanu, A., and Monnier, L. (1997). Nonfasting plasma glucose is a better marker of diabetic control than fasting plasma glucose in type 2 diabetes. *Diabetes Care*, 20:1822–1826.

Belitz, C., Brezger, A., Klein, N., Kneib, T., Lang, S., and Umlauf, N. (2015). BayesX: Software for Bayesian inference in structured additive regression models. Version 3.0.2. *Available online on http://www.BayesX.org/*.

Belitz, C., Brezger, A., Kneib, T., Lang, S., and Umlauf, N. (2016). BayesX: Software for Bayesian inference in structured additive regression models. Version 1.1. *Available on http://www.BayesX.org/*.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, 37:905–938.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43:1–20.

Bonora, E., Calcaterra, F., Lombardi, S., Bonfante, N., Formentini, G., Bonadonna, R. C., and Muggeo, M. (2001). Plasma glucose levels throughout the day and HbA1c interrelationships in type 2 diabetes: Implications for treatment and monitoring of metabolic control. *Diabetes Care*, 24:2023–2029.

Bowman, A. W. and Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*. New York, USA: Oxford University Press, Inc.

Breiman, L. (1998). Arcing classifier (with discussion and a rejoinder by the author). *The Annals of Statistics*, 26:801–849.

Brezger, A. and Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis*, 50:967–991.

Brockhaus, S., Fuest, A., Mayr, A., and Greven, S. (2018). Signal regression models for location, scale and shape with an application to stock returns. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67:665–686.

Brockhaus, S. and Ruegamer, D. (2018). FDboost: Boosting functional regression models. R package version 0.3-1. *Available on CRAN*.

Brockhaus, S., Ruegamer, D., and Greven, S. (2017). Boosting functional regression models with FDboost. *ArXiv e-prints*.

Cardot, H. (2002). Spatially adaptive splines for statistical linear inverse problems. *Journal of Multivariate Analysis*, 81:100–119.

Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics and Probability Letters*, 45:11–22.

Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13:571–591.

Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, 92:24–41.

Cohen, R. M., Holmes, Y. R., Chenier, T. C., and Joiner, C. H. (2003). Discordance between HbA1c and fructosamine: Evidence for a glycosylation gap and its relation to diabetic nephropathy. *Diabetes Care*, 26:163–167.

Cole, T. J. and Green, P. J. (1992). Smoothing reference centile curves: The lms method and penalized likelihood. *Statistics in Medicine*, 11:1305–1319.

Craig, C., Marshall, A., Sjorstrom, M., Bauman, A., Booth, M., Ainsworth, B., Pratt, M., Ekelund, U., Yngve, A., and Sallis, J. (2003). International physical activity questionnaire: 12-country reliability and validity. *Medicine & Science in Sports & Exercise*, 35:1381–1395.

Cuevas, A., Febrero, M., and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22:481–496.

Davidson, M. B. (1979). The effect of aging on carbohydrate metabolism: A review of the English literature and a practical approach to the diagnosis of diabetes mellitus in the elderly. *Metabolism-Clinical and Experimental*, 28:688–705.

Dierckx, P. (1995). *Curve and surface fitting with splines*. Oxford, United Kingdom: Oxford University Press.

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5:236–244.

Durbán, M. (2009). An introduction to smoothing with penalties: P-splines. *Boletín de Estadística e Investigación Operativa*, 25:195–205.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89–121.

Escabias, M., Aguilera, A. M., and Valderrama, M. J. (2004). Principal component estimation of functional logistic regression: Discussion of two different approaches. *Journal of Nonparametric Statistics*, 16:365–384.

Espasandín-Domínguez, J., Benítez-Estévez, A. J., Cadarso-Suárez, C., Kneib, T., Barreiro-Martínez, T., Casas-Méndez, B., and Gude, F. (2018a). Geographical differences in blood potassium detected using a structured additive distributional regression model. *Spatial Statistics*, 24:1–13.

Espasandín-Domínguez, J., Carollo-Limeres, C., Coladas-Uria, L., Cadarso-Suárez, C., Lado-Baleato, O., and Gude, F. (2018b). Bivariate copula additive models for location, scale and shape with applications in biomedicine. In *Gil E., Gil E., Gil J., Gil M. (eds) The Mathematics of the Uncertaing*, volume 142, pages 135–146. Verlag: Springer International Publishing.

Fahrmeir, L. and Kneib, T. (2011). *Bayesian smoothing and regression for longitudinal, spatial and event history data*. Oxford, United Kingdom: Oxford Statistical Science Series.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression. Models, methods and applications.* Heidelberg, Germany: Springer.

Febrero-Bande, M., Galeano, P., and González-Manteiga, W. (2017). Functional principal component regression and functional partial least-squares regression: An overview and a comparative study. *International Statistical Review*, 85:61–83.

Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51:1–28.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: Theory and practice*. New York, USA: Springer Science & Business Media, Inc.

Filippou, P., Marra, G., and Radice, R. (2017). Penalized likelihood estimation of a trivariate probit model. *Biostatistics*, 18:569–585.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, ICML'96, pages 148–156. Morgan Kaufmann Publishers, Inc.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29:1189–1232.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823.

Gijbels, I., Omelka, M., and Veraverbeke, N. (2012). Multivariate and functional covariates and conditional copulas. *Electronic Journal of Statistics*, 6:1273–1306.

Gijbels, I., Veraverbeke, N., and Omelka, M. (2011). Conditional copulas, association measures and their applications. *Computational Statistics & Data Analysis*, 55:1919–1932.

Green, L. (2006). *Queueing analysis in healthcare*. New York, USA: Springer and Business Media, LLC.

Greven, S. and Scheipl, F. (2017). A general framework for functional regression modelling. *Statistical Modelling*, 17:1–35.

Gruen, M. E., Alfaro-Córdoba, M., Thomson, A. E., Worth, A. C., Staicu, A. M., and Lascelles, D. B. (2017). The use of functional data analysis to evaluate activity in a spontaneous model of degenerative joint disease associated pain in cats. *PloS ONE*, 12:e0169576.

Gual, A., Martos, A. R., Lligoña, A., and Llopis, J. (1999). Does the concept of a standard drink apply to viticultural societies? *Alcohol and Alcoholism (Oxford, Oxfordshire)*, 34:153–160.

Gude, F., Díaz-Vidal, P., Rúa-Pérez, C., Alonso-Sampedro, M., Fernández-Merino, C., Rey-García, J., Cadarso-Suárez, C., Pazos-Couselo, M., García-López, J., and González-Quintela, A. (2017). Glycemic variability and its association with demographics and lifestyles in a general adult population. *Journal of Diabetes Science and Technology*, 11:780–790.

Hall, P., Müller, H. G., and Wang, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 34:1493–1517.

Hallworth, M. J. (2011). *The "70% claim": What is the evidence base?* London, England: SAGE Publications Sage UK.

Hashimoto, Y., Futamura, A., and Ikushima, M. (1995). Effect of aging on HbA1c in a working male japanese population. *Diabetes Care*, 18:1337–1340.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. Boca Raton, USA: Chapman and Hall/CRC Monographs on Statistics and Applied Probability.

Hofner, B., Mayr, A., Robinzonov, N., and Schmid, M. (2014). Model-based boosting in R: A hands-on tutorial using the R package mboost. *Computational Statistics*, 29:3–35.

Horowitz, G. L. (2008). *Defining, establishing, and verifying reference intervals in the clinical laboratory: Proposed guideline*. Clinical and Laboratory Standards Institute: CLSI Guidelines.

Huh, J. H., Kim, K. J., Lee, B.-W., Kim, D. W., Kang, E. S., Cha, B. S., and Lee, H. C. (2014). The relationship between BMI and glycated albumin to glycated hemoglobin (GA/A1c) ratio according to glucose tolerance status. *PloS ONE*, 9:e89478.

Instituto Galego de Estatística (2015). Indicadores de seguimento das directrices de ordenación do territorio. *Available online on http://www.ige.eu/igebdt/igeapi/datos/77/0:1996,9915:12*.

James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63:533–550.

James, G. M. and Silverman, B. W. (2005). Functional adaptive model estimation. *Journal of the American Statistical Association*, 100:565–576.

Jeffreys, H. (1998). *The theory of probability*. Oxford, United Kingdom: Oxford University Press.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. London, New York: Chapman and Hall/CRC.

Juraschek, S., Steffes, M., and Selvin, E. (2012). Associations of alternative markers of glycemia with hemoglobin a(1c) and fasting glucose. *Clinical Chemistry*, 58:1648–1655.

Kammann, E. and Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52:1–18.

Kendrick, S. K., Zheng, Q., Garbett, N. C., and Brock, G. N. (2017). Application and interpretation of functional data analysis techniques to differential scanning calorimetry data from lupus patients. *PloS ONE*, 12:e0186232.

Klein, N. and Kneib, T. (2016a). Scale-dependent priors for variance parameters in structured additive distributional regression. *Bayesian Analysis*, 11:1071–1106.

Klein, N. and Kneib, T. (2016b). Simultaneous inference in structured additive conditional copula regression models: A unifying Bayesian approach. *Statistics and Computing*, 26:841–860.

Klein, N., Kneib, T., Klasen, S., and Lang, S. (2014a). Bayesian structured additive distributional regression for multivariate responses. *Journal of the Royal Statistical Society: Series C*, 64:569–591.

Klein, N., Kneib, T., Klasen, S., and Lang, S. (2015). Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *The Annals of Applied Statistics*, 9:1024–1052.

Klein, N., Kneib, T., and Lang, S. (2014b). Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data. *Journal of the American Statistical Association*, 110:405–419.

Kneib, T., Heinzl, F., Brezger, A., Bové, D. S., and Klein, N. (2014). BayesX: R utilities accompanying the software package BayesX. R package version 0.2-9. *Available on CRAN*.

Koenig, R. J., Peterson, C. M., Jones, R. L., Saudek, C., Lehrman, M., and Cerami, A. (1976). Correlation of glucose regulation and hemoglobin AIc in diabetes mellitus. *New England Journal of Medicine*, 295:417–420.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 46:33–50.

Koenker, R. and Ng, P. (2005). Inequality constrained quantile regression. *Sankhyā: The Indian Journal of Statistics*, 67:418–440.

Koga, M., Matsumoto, S., Saito, H., and Kasayama, S. (2006). Body mass index negatively influences glycated albumin, but not glycated hemoglobin, in diabetic patients. *Endocrine Journal*, 53:387–391.

Krämer, N., Boulesteix, A. L., and Tutz, G. (2008). Penalized partial least squares with applications to B-spline transformations and functional data. *Chemometrics and Intelligent Laboratory Systems*, 94:60–69.

Lang, S. and Brezger, A. (2004). Bayesian P-spline. *Journal of Computational and Graphical Statistics*, 13:183–212.

Li, H., Xiao, G., Xia, T., Tang, Y. Y., and Li, L. (2014). Hyperspectral image classification using functional data analysis. *IEEE Transactions on Cybernetics*, 44:1544–1555.

López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104:718–734.

Mankowska, D., Meisel, F., and Bierwirth, C. (2014). The home health care routing and scheduling problem with interdependent services. *Health Care Management Science*, 17:15–30.

Marra, G. and Radice, R. (2017a). Bivariate copula additive models for location, scale and shape. *Computational Statistics and Data Analysis*, 112:99–113.

Marra, G. and Radice, R. (2017b). GJRM: Generalised joint regression modelling. R package version 0.1-2. *Available on CRAN*.

Mayr, A., Fenske, N., Hofner, B., Kneib, T., and Schmid, M. (2012a). Generalized additive models for location, scale and shape for high dimensional data-a flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61:403–427.

Mayr, A., Hofner, B., and Schmid, M. (2012b). The importance of knowing when to stop. *Methods of Information in Medicine*, 51:178–186.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. London, New York: Chapman and Hall/CRC Monographs on Statistics and Applied Probability.

McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F., and Ruppert, D. (2014). Functional generalized additive models. *Journal of Computational and Graphical Statistics*, 23:249–269.

Miyashita, Y., Nishimura, R., Morimoto, A., Matsudaira, T., Sano, H., and Tajima, N. (2007). Glycated albumin is low in obese, type 2 diabetic patients. *Diabetes Research and Clinical Practice*, 78:51–55.

Monnier, L. and Colette, C. (2009). Target for glycemic control: Concentrating on glucose. *Diabetes Care*, 32:S199–S204.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359.

Müller, H. G. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, 33:774–805.

Müller, H. G. and Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association*, 103:1534–1544.

Nakashima, K., Nishizaki, O., and Andoh, Y. (1993). Acceleration of hemoglobin glycation with aging. *Clinica Chimica Acta*, 215:111–118.

Nathan, D., Turgeon, H., and Regan, S. (2007). Relationship between glycated haemoglobin levels and mean glucose levels over time. *Diabetologia*, 50:2239–2244.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135:370–384.

Nelsen, R. (2006). *An introduction to copulas*. Heidelberg, Germany: Springer-Verlag.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, 55:819–847.

Palaro, H. P. and Hotta, L. K. (2006). Using conditional copula to estimate value at risk. *Journal of Data Science*, 4:93–115.

Palmer, B. F. and Clegg, D. J. (2016). Physiology and pathophysiology of potassium homeostasis. *Advances in Physiology Education*, 40:480–490.

Pani, L. N., Korenda, L., Meigs, J. B., Driver, C., Chamany, S., Fox, C. S., Sullivan, L., D'Agostino, R. B., and Nathan, D. M. (2008). Effect of aging on A1C levels in individuals without diabetes: Evidence from the Framingham Offspring Study and the National Health and Nutrition Examination Survey 2001–2004. *Diabetes Care*, 31:1991–1996.

Preda, C. and Saporta, G. (2005). Clusterwise PLS regression on a stochastic process. *Computational Statistics & Data Analysis*, 49:99–108.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. *Available online on https://www.R-project.org/*.

Radice, R., Marra, G., and Wojtyś, M. (2016). Copula regression spline models for binary outcomes. *Statistics and Computing*, 26:981–995.

Ramsay, J. and Silverman, B. (1997). *The analysis of functional data*. Berlin, Germany: Springer-Verlag.

Ramsay, J. O. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53:539–572.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. New York, USA: Springer Science & Business Media, Inc.

Ramsay, J. O. and Silverman, B. W. (2007). *Applied functional data analysis: Methods and case studies*. New York: Springer.

Reiss, P. T. and Ogden, R. T. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102:984–996.

Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57:253–259.

Rigby, A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics*, 54:507–554.

Rigby, R. A. and Stasinopoulos, D. M. (1996). Mean and dispersion additive models. In *Statistical theory and computational aspects of smoothing*, volume 104, pages 215–230. Semmering, Austria: Springer.

Rohlfing, C. L., Wiedmeyer, H.-M., Little, R. R., England, J. D., Tennill, A., and Goldstein, D. E. (2002). Defining the relationship between plasma glucose and HbA1c: Analysis of glucose profiles and HbA1c in the diabetes control and complications trial. *Diabetes Care*, 25:275–278.

Rossi, F., Delannay, N., Conan-Guez, B., and Verleysen, M. (2005). Representation of functional data in neural networks. *Neurocomputing*, 64:183–210.

Sabeti, A., Wei, M., and Craiu, R. V. (2014). Additive models for conditional copulas. *Statistics*, 3:300–312.

Sacks, D. (2011). A1c versus glucose testing: A comparison. *Diabetes Care*, 34:518–523.

Sakuma, N., Omura, M., Oda, E., and Saito, T. (2011). Converse contributions of fasting and postprandial glucose to hba 1c and glycated albumin. *Diabetology International*, 2:162–171.

Scheipl, F., Gertheiss, J., and Greven, S. (2016). Generalized functional additive mixed models. *Electronic Journal of Statistics*, 10:1455–1492.

Schnabel, S. K. and Eilers, P. H. (2009). An analysis of life expectancy and economic production using expectile frontier zones. *Demographic Research*, 21:109–134.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris*, 8:229–231.

Sobotka, F. and Kneib, T. (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis*, 56:755–767.

Spiegelhalter, D., Best, N., Carlin, B., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:583–639.

Stankovic, A. and Smith, S. (2004). Elevated serum potassium values. The role of preanalytic variables. *American Journal of Clinical Pathology*, 121:105–112.

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23:1–46.

Tanner, M., Kent, N., Smith, B., Fletcher, S., and Lewer, M. (2008). Stability of common biochemical analytes in serum gel tubes subjected to various storage temperatures and times pre-centrifugation. *Annals of Clinical Biochemistry*, 121:375–379.

The Diabetes Control and Complications Trial Research Group (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *The New England Journal of Medicine*, 329:977–986.

Tian, T. S. (2010). Functional data analysis in brain imaging studies. *Frontiers in Psychology*, 1:35–45.

U.K. Prospective Diabetes Study (UKPDS) Group (1998a). Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (ukpds 34). *The Lancet*, 352:854–865.

U.K. Prospective Diabetes Study (UKPDS) Group (1998b). Intensive blood glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (ukpds 33). *The Lancet*, 352:837–853.

Umlauf, N., Adler, D., Kneib, T., Lang, S., and Zeileis, A. (2015). Structured additive regression models: An R interface to BayesX. *Journal of Statistical Software*, 63:1–46.

Umlauf, N., Kneib, T., and Klein, N. (2018). BayesX: R utilities accompanying the software package BayesX. R package version 0.3-0. *Available on CRAN*.

Vatter, T. and Chavez-Demoulin, V. (2015). Generalized additive models for conditional dependence structures. *Journal of Multivariate Analysis*, 141:147–167.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14:867–897.

Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 121:375–379.

Wood, S. (2006). *Generalized additive models: An introduction with R*. Boca Raton, USA: Chapman and Hall.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:3–36.

Wood, S. (2017). *Generalized additive models: An introduction with R*. Boca Raton, USA: Chapman and Hall/CRC Texts in Statistical Science.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111:1548–1563.

Yan, J. (2007). Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software*, 21:1–21.

Yao, F. and Lee, T. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68:3–25.

Yee, T. W. (2015). *Vector generalized linear and additive models: With an implementation in R*. New York, USA: Springer.

Yee, T. W. and Wild, C. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:481–493.

Zhou, J., Li, H., Ran, X., Yang, W., Li, Q., Peng, Y., Li, Y., Gao, X., Luan, X., and Wang, W. (2009). Reference values for continuous glucose monitoring in Chinese subjects. *Diabetes Care*, 32:1188–1193.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Appendix A

Supplementarial material to Chapter 3: "Detecting differences in blood potassium concentrations by using a spatial distributional regression model"

A.1 Software and code

Statistical analyses of Chapter 3 were performed using open-source BayesX software (Belitz et al., 2015). The BayesX (Umlauf et al., 2018) and R2BayesX (Umlauf et al., 2015; Belitz et al., 2016) R packages were used as graphic interfaces.

In the following, we specified the code needed to estimate a distributional regression model considering a log-normal response and spatial information using BayesX software.

```
%-------------------------------------------------------------
% Reading dataset information %
%-------------------------------------------------------------

%usefile E:\Potassium\Lognormal.txt

dataset d
d.infile using E:\Potassium\Potassium.txt
```

```
%-------------------------------------------------------------
% Map objects %
%-------------------------------------------------------------

map m
m.infile, graph using E:\Potassium\Santiago.gra


%-------------------------------------------------------------
% Bayesian  distributional  regression %
%-------------------------------------------------------------
mcmcreg r
r.outfile = E:\Potassium\log
logopen, replace using E:\Potassium\logreml.txt

r.hregress k=const + age(pspline) + time(pspline)+sex +
 CP(spatial, map=m)+ CP(random),
 iterations=9000 burnin=2000 step=10 family=lognormal
 equationtype=sigma2 using d

r.hregress k=const + age(pspline) +time(pspline)+ sex +
 CP(spatial, map=m)+ CP(random),
 family=lognormal predict=full equationtype=mu
 setseed=6 using d


logclose
```

Once BayesX model is adjusted. Several files have been generated. These files are necessary to plot the results. For example, to plot the effect of age on mean and variance of potassium levels, the following code is needed:

```
%-------------------------------------------------------------
#  \beta_mu (Age)
%-------------------------------------------------------------
a=read.table("lognormal_MAIN_mu_REGRESSION_k_nonlinear_
pspline_effect_of_age.res",header=T)
plot2d(a$pqu2p5 + a$pmean + a$pqu97p5 ~ a$age, main="",
 xlab="Age (Years)",
 ylab="Effect of Age",fill.select = c(0, 1, 0, 1), lty = c(0,
 1, 0),col="black",
 col.lines = "deeppink4",lwd.lines=2,  col.residuals="red",
 col.polygons = "pink")
```

```
grid(col="grey") # grid only in y-directionn
abline(h=0,lty=2,col="blue")

%-------------------------------------------------------------
#  \beta_sigma (Age)
%-------------------------------------------------------------

a=read.table("lognormal_MAIN_sigma2_REGRESSION_k_nonlinear_
pspline_effect_of_age.res",header=T)
plot2d(a$pqu2p5^2 + a$pmean + a$pqu97p5 ~ a$age,main="",
 xlab="Age (Years)",
 ylab="Effect of Age",fill.select = c(0, 1, 0, 1), lty = c(0,
 1, 0),col="black",col.lines = "lightsteelblue4",lwd.lines=2,
 col.residuals="red",col.polygons = "lightsteelblue2")
grid(col="grey") # grid only in y-direction
abline(h=0,lty=2,col="deeppink4")
```

A.2 Visualization of the obtained results

In the following, we specified the code created to visualize the results - on the real scale - of a distributional regression models comprising spatial information of the response for potassium dataset. Note that once the model is adjusted in BayesX, the representation of the effect can be made using R software.

```
# ------------------------------------------------------------
# RESULTS - VISUALIZATION IN R ---------------------------
# ------------------------------------------------------------

library("splines")

data <- read.table("Potassium.txt", header=TRUE)
data2 <- read.table("Average.raw", header=TRUE)

# data2 is a dataset which includes the variables of the com-
plete "potassium dataset", together with the mean values of
age, time and potassium for each district.

# Effects on mu and sigma2:
# - const
# - sex
# - age (pspline)
# - time (pspline)
```

```
# - CP (random)
# - CP (MRF)
# read all samples
# ------------------------------------------------------------
age_mu_samples <-
 read.table("logn_MAIN_mu_REGRESSION_k_nonlinear_pspline_
 effect_of_age_sample.raw", header=TRUE)[,-1]
age_sigma2_samples <- read.table("logn_MAIN_sigma2_REGRESSION
 _k_ nonlinear_pspline_effect_of_age_ sample.raw",header=
 TRUE)[,-1]
time_mu_samples <- read.table("logn_MAIN_mu_REGRESSION_k_
 nonlinear_pspline_effect_of_time_sample.raw", header=
 TRUE)[,-1]
 time_sigma2_samples <-read.table("logn_MAIN_sigma2_REGRESSION
 _k_nonlinear_pspline_effect_of_time_ sample.raw", header=
 TRUE)[,-1]
fixed_mu_samples <- read.table("logn_MAIN_mu_REGRESSION_k_
fixed_sigma2_samples <- read.table("logn_MAIN_sigma2_
 REGRESSION_k_Lineareffects_sample.raw",  header=
 TRUE)[,-1,drop=FALSE]
CP_random_mu_samples <- read.table("logn_MAIN_mu_REGRESSION_k_
 random_effect_of_CP_sample.raw", header=TRUE)[,-1,drop=FALSE]
CP_random_sigma2_samples <- read.table("logn_MAIN_sigma2_
 REGRESSION_k_random_effect_of_CP_sample.raw", header=TRUE)
 [,-1,drop=FALSE]
CP_mrf_mu_samples <- read.table("logn_MAIN_mu_REGRESSION_k_
 spatial_MRF_effect_of_CP_sample.raw", header=TRUE)
 [,-1,drop=FALSE]
CP_mrf_sigma2_samples <- read.table("logn_MAIN_sigma2_
 REGRESSION_k_spatial_MRF_effect_of_CP_sample.raw",header
 =TRUE)[,-1,drop=FALSE]

# ------------------------------------------------------------
# Visualise effect of age for given values of all other
# covariates:
# ------------------------------------------------------------
# construct a design matrix for age with 100 equidistant
# age values:

age_step <- (max(data$age)-min(data$age))/19+0.01
age_knots <- seq(min(data$age)-3*age_step, max(data$age)+3*
age_step, length=20+2*3)
age_seq <- seq(min(data$age), max(data$age), length=100)
```

```
age_B <- spline.des(age_knots, age_seq, 4)$design

# construct a design matrix for time with the average value
# of time
time_step <-(max(data$time)-min(data$time))/19+0.01
time_knots <-seq(min(data$time)-3*time_step, max(data$time)+
 3*time_step, length=20+2*3)
time_seq <-rep(mean(unique(data$time)), 100)
time_B <-spline.des(time_knots, time_seq, 4)$design

# ------------------------------------------------------------
# construct a design matrix for the intercept and for sex
# ------------------------------------------------------------
fixed_B <- cbind(rep(1,100), rep(2,100))

eta_mu_age <- matrix(0, nrow=100, ncol=700)
eta_sigma2_age <- matrix(0, nrow=100, ncol=700)
for(i in 1:700)
{
eta_mu_age[,i] <- fixed_B %*% t(fixed_mu_samples[i,])+
 age_B %*% t(age_mu_samples[i,]) +
time_B %*% t(time_mu_samples[i,])
 eta_sigma2_age[,i] <- fixed_B %*% t(fixed_sigma2_
 samples[i,])+ age_B %*% t(age_sigma2_samples[i,])+
 time_B %*% t(time_sigma2_samples[i,])
}

# Now the effect of time

age_step <- (max(data$age)-min(data$age))/19+0.01
age_knots <- seq(min(data$age)-3*age_step, max(data$age)+
 3*age_step, length=20+2*3)
age_seq <- rep(mean(unique(data$age)), 100)
age_B <- spline.des(age_knots, age_seq, 4)$design

# ------------------------------------------------------------
# construct a design matrix for time with the average value
# of time
# ------------------------------------------------------------
time_step <- (max(data$time)-min(data$time))/19+0.01
time_knots <-seq(min(data$time)-3*time_step, max(data$time)+
 3*time_step, length=20+2*3)
time_seq <- seq(min(data$time), max(data$time), length=100)
```

```
time_B <- spline.des(time_knots, time_seq, 4)$design

eta_mu_time <- matrix(0, nrow=100, ncol=700)
eta_sigma2_time <- matrix(0, nrow=100, ncol=700)
for(i in 1:700)
{
eta_mu_time[,i] <- fixed_B %*% t(fixed_mu_samples[i,])+
 age_B %*% t(age_mu_samples[i,]) +
 time_B %*% t(time_mu_samples[i,])
eta_sigma2_time[,i]<-fixed_B %*%t(fixed_sigma2_samples[i,])+
 age_B %*% t(age_sigma2_samples[i,]) +
 time_B %*% t(time_sigma2_samples[i,])
}

# ------------------------------------------------------------
# Now the spatial effects
# ------------------------------------------------------------

CP_seq_mrf <- read.table("logn_MAIN_mu_REGRESSION_k_spatial_
MRF_ effect_of_CP.res", header=TRUE)[,2]
o <- order(CP_seq_mrf)
CP_mrf_mu_samples <- CP_mrf_mu_samples[,o]
CP_mrf_sigma2_samples <- CP_mrf_sigma2_samples[,o]
CP_seq_random <- read.table("logn_MAIN_mu_REGRESSION_k_
 random_effect_of_CP.res", header=TRUE)[,2]

data2 <- data2[!duplicated(data2$CP),]
data2 <- data2[order(data2$CP),]

age_step <- (max(data$age)-min(data$age))/19+0.01
age_knots <- seq(min(data$age)-3*age_step, max(data$age)+
 3*age_step, length=20+2*3)
#age_seq <- rep(mean(unique(data$age)), 46)
age_seq <- data2$age_mean
age_B <- spline.des(age_knots, age_seq, 4)$design

time_step <- (max(data$time)-min(data$time))/19+0.01
time_knots <- seq(min(data$time)-3*time_step,
 max(data$time)+ 3*time_step, length=20+2*3)
#time_seq <- rep(mean(unique(data$age)), 46)
time_seq <- data2$time_mean
time_B <- spline.des(time_knots, time_seq, 4)$design
```

```
fixed_B <- cbind(rep(1,46), rep(2,46))

eta_mu_mrf <- matrix(0, nrow=46, ncol=700)
eta_sigma2_mrf <- matrix(0, nrow=46, ncol=700)
eta_mu_random <- matrix(0, nrow=46, ncol=700)
eta_sigma2_random <- matrix(0, nrow=46, ncol=700)
eta_mu_total <- matrix(0, nrow=46, ncol=700)
eta_sigma2_total <- matrix(0, nrow=46, ncol=700)

for(i in 1:700)
{
eta_mu_random[,i] <- fixed_B %*% t(fixed_mu_samples[i,])+
 age_B %*% t(age_mu_samples[i,]) +
 time_B %*% t(time_mu_samples[i,]) +
 t(CP_random_mu_samples[i,])
eta_sigma2_random[,i]
 <-fixed_B %*%t(fixed_sigma2_samples[i,])+
 age_B %*% t(age_sigma2_samples[i,]) +
 time_B %*% t(time_sigma2_samples[i,]) +
 t(CP_random_sigma2_samples[i,])
eta_mu_mrf[,i] <- fixed_B %*% t(fixed_mu_samples[i,])+
 age_B %*% t(age_mu_samples[i,]) +
 time_B %*% t(time_mu_samples[i,]) +
 t(CP_mrf_mu_samples[i,])
eta_sigma2_mrf[,i]<-fixed_B %*%t(fixed_sigma2_samples[i,])+
 age_B %*% t(age_sigma2_samples[i,]) +
 time_B %*% t(time_sigma2_samples[i,]) +
 t(CP_mrf_sigma2_samples[i,])
eta_mu_total[,i] <-
 eta_mu_random[,i] + t(CP_mrf_mu_samples[i,])
eta_sigma2_total[,i] <-
 eta_sigma2_random[,i] + t(CP_mrf_sigma2_samples[i,])
}

# functions to compute the mean and the variance of
# the log-normal for given predictors

lnmean <- function(eta_mu, eta_sigma2)
{
exp(eta_mu+0.5*exp(eta_sigma2))
}
lnvar <- function(eta_mu, eta_sigma2)
{
```

```
(exp(exp(eta_sigma2))-1)*(exp(2*eta_mu+exp(eta_sigma2)))
}
# ------------------------------------------------------------
# transform the samples from the predictor level to the mean
# and variance
# ------------------------------------------------------------
mean_age <- lnmean(eta_mu_age, eta_sigma2_age)
mean_time <- lnmean(eta_mu_time, eta_sigma2_time)
mean_mrf <- lnmean(eta_mu_mrf, eta_sigma2_mrf)
mean_random <- lnmean(eta_mu_random, eta_sigma2_random)
mean_total <- lnmean(eta_mu_total, eta_sigma2_total)

var_age <- lnvar(eta_mu_age, eta_sigma2_age)
var_time <- lnvar(eta_mu_time, eta_sigma2_time)
var_mrf <- lnvar(eta_mu_mrf, eta_sigma2_mrf)
var_random <- lnvar(eta_mu_random, eta_sigma2_random)
var_total <- lnvar(eta_mu_total, eta_sigma2_total)

# computation of the posterior mean and credible intervals
# for the effects
mean_age_mean <- apply(mean_age, 1, mean)
mean_age_q2p5 <- apply(mean_age, 1, quantile, prob=0.025)
mean_age_q97p5 <- apply(mean_age, 1, quantile, prob=0.975)

mean_time_mean <- apply(mean_time, 1, mean)
mean_time_q2p5 <- apply(mean_time, 1, quantile, prob=0.025)
mean_time_q97p5 <-apply(mean_time, 1, quantile,prob=0.975)

mean_mrf_mean <- apply(mean_mrf, 1, mean)
mean_mrf_q2p5 <- apply(mean_mrf, 1, quantile, prob=0.025)
mean_mrf_q97p5 <- apply(mean_mrf, 1, quantile, prob=0.975)
mean_mrf_p95 <- -1*(mean_mrf_q97p5<mean(mean_mrf_mean))
 +1*(mean_mrf_q2p5>mean(mean_mrf_mean))

mean_random_mean <- apply(mean_random, 1, mean)
mean_random_q2p5 <- apply(mean_random, 1, quantile,
 prob=0.025)
mean_random_q97p5 <- apply(mean_random, 1, quantile,
prob=0.975)
mean_random_p95 <- -1*(mean_random_q97p5<
mean(mean_random_mean))+ 1*(mean_random_q2p5>
mean(mean_random_mean))
```

```
mean_total_mean <- apply(mean_total, 1, mean)
mean_total_q2p5 <- apply(mean_total, 1, quantile, prob=
 0.025)
mean_total_q97p5 <- apply(mean_total, 1, quantile, prob=
 0.975)
mean_total_p95<- -(mean_total_q97p5<mean(mean_total_mean))
 + 1*(mean_total_q2p5>mean(mean_total_mean))

var_age_mean <- apply(var_age, 1, mean)
var_age_q2p5 <- apply(var_age, 1, quantile, prob=0.025)
var_age_q97p5 <- apply(var_age, 1, quantile, prob=0.975)

var_time_mean <- apply(var_time, 1, mean)
var_time_q2p5 <- apply(var_time, 1, quantile, prob=0.025)
var_time_q97p5 <- apply(var_time, 1, quantile, prob=0.975)

var_mrf_mean <- apply(var_mrf, 1, mean)
var_mrf_q2p5 <- apply(var_mrf, 1, quantile, prob=0.025)
var_mrf_q97p5 <- apply(var_mrf, 1, quantile, prob=0.975)
var_mrf_p95 <- -1*(var_mrf_q97p5<mean(var_mrf_mean)) +
 1*(var_mrf_q2p5>mean(var_mrf_mean))

var_random_mean <- apply(var_random, 1, mean)
var_random_q2p5 <- apply(var_random, 1, quantile, prob=
 0.025)
var_random_q97p5 <- apply(var_random, 1, quantile, prob=
 0.975)
var_random_p95 <- -(var_random_q97p5<mean(var_random_mean))
 + 1*(var_random_q2p5>mean(var_random_mean))

var_total_mean <- apply(var_total, 1, mean)
var_total_q2p5 <- apply(var_total, 1, quantile, prob=0.025)
var_total_q97p5 <- apply(var_total, 1, quantile, prob=0.975)
var_total_p95 <- -1*(var_total_q97p5<mean(var_total_mean))
 + 1*(var_total_q2p5>mean(var_total_mean))

# ------------------------------------------------------------
# produce the plots
# ------------------------------------------------------------
library("BayesX")
m <- read.bnd("Santiago.bnd") # read the map
m2 <- read.gra("Santiago.gra")
age_seq <- seq(min(data$age), max(data$age), length=100)
```

```
time_seq <- seq(min(data$time), max(data$time), length=100)
plotspatial <- data.frame(CP_seq=CP_seq_random,
mean_mrf_mean, mean_random_mean, mean_total_mean,
mean_mrf_p95, mean_random_p95, mean_total_p95,
var_mrf_mean, var_random_mean, var_total_mean,
var_mrf_p95, var_random_p95, var_total_p95)

# ------------------------------------------------------------
# effects on the mean
# ------------------------------------------------------------
par(mfrow=c(2,4))
plot(age_seq, mean_age_mean, type="l", xlab="Age", ylab="",
ylim=c(min(mean_age_q2p5)-0.1, max(mean_age_q97p5)+0.1))
lines(age_seq, mean_age_q2p5, lty=2,col="pink")
lines(age_seq, mean_age_q97p5, lty=2,col="pink")
polygon(c(age_seq,rev(age_seq)),c(mean_age_q97p5,
 rev(mean_age_q2p5)),col="pink",border=NA)
lines(age_seq, mean_age_mean, type="l",
 col = "deeppink4",lwd=2)
grid(NA, 5,col="grey")
grid(5, NA, lwd = 1,col="grey")
abline(h=0,lty=2,lwd=2)
rug(data$age,lwd=2)

par(mfrow=c(1,3))

drawmap(data=plotspatial, map=m, regionvar="CP_seq",
 plotvar="mean_mrf_mean")
drawmap(data=plotspatial, map=m, regionvar="CP_seq",
 plotvar="mean_random_mean")
drawmap(data=plotspatial, map=m, regionvar="CP_seq",
 plotvar="mean_total_mean")

plot(time_seq, mean_time_mean, type="l", xlab="cctime",
 ylim=c(min(mean_time_q2p5), max(mean_time_q97p5)))
lines(time_seq, mean_time_q2p5, lty=2,col="pink")
lines(time_seq, mean_time_q97p5, lty=2,col="pink")
polygon(c(time_seq,rev(time_seq)),c(mean_time_q97p5,
 rev(mean_time_q2p5)),  col="pink",border=NA)
lines(time_seq, mean_time_mean, type="l",
 col = "deeppink4",lwd=2)
grid(NA, 5,col="grey")
grid(5, NA, lwd = 1,col="grey")
```

```
abline(h=0,lty=2,lwd=2)
rug(data$time,lwd=2)
axis(4,labels=F,tick=T,col.ticks="white")
axis(2,labels=F,tick=T)

drawmap(data=plotspatial, map=m, regionvar="CP_seq",
 plotvar="mean_mrf_p95", pcat=TRUE)
drawmap(data=plotspatial, map=m, regionvar="CP_seq",
  plotvar="mean_random_p95", pcat=TRUE)
drawmap(data=plotspatial, map=m, regionvar="CP_seq",
 plotvar="mean_total_p95", pcat=TRUE)

# ------------------------------------------------------------
# effects on the variance
# ------------------------------------------------------------

par(mfrow=c(1,3))
plot(age_seq, var_age_mean, type="l", xlab="Age", ylab="",
ylim=c(min(var_age_q2p5)-0.02, max(var_age_q97p5)))
lines(age_seq, var_age_q2p5, lty=2,col="lightsteelblue2")
lines(age_seq, var_age_q97p5, lty=2,col="lightsteelblue2")
polygon(c(age_seq,rev(age_seq)),c(var_age_q97p5,
rev(var_age_q2p5)),col="lightsteelblue2",border=NA)
lines(age_seq, var_age_mean, type="l",
col = "lightsteelblue4",lwd=2)
grid(NA, 5,col="grey")
grid(5, NA, lwd = 1,col="grey")
rug(data$age,lwd=2)


drawmap(data=plotspatial, map=m, regionvar="CP_seq",
 plotvar="var_mrf_mean")
drawmap(data=plotspatial, map=m, regionvar="CP_seq",
 plotvar="var_random_mean")
drawmap(data=plotspatial, map=m, regionvar="CP_seq",
 plotvar="var_total_mean")

plot(time_seq, var_time_mean, type="l", xlab="cctime",
 ylab="variance",
 ylim=c(min(var_time_q2p5), max(var_time_q97p5)))
lines(time_seq, var_time_q2p5, lty=2,
col="lightsteelblue2")
lines(time_seq, var_time_q97p5, lty=2,
```

```
col="lightsteelblue2")
polygon(c(time_seq,rev(time_seq)),c(var_time_q97p5,
 rev(var_time_q2p5)), col="lightsteelblue2",border=NA)
lines(time_seq, var_time_mean, type="l",
 col = "lightsteelblue4",lwd=2)
grid(NA, 5,col="grey")
grid(5, NA, lwd = 1,col="grey")
abline(h=0,lty=2,lwd=2)
rug(data$time,lwd=2)
axis(4,labels=F,tick=T,col.ticks="white")
axis(2,labels=F,tick=T)

drawmap(data=plotspatial, map=m, regionvar="CP_seq",
 plotvar="var_mrf_p95", pcat=TRUE)
drawmap(data=plotspatial, map=m, regionvar="CP_seq",
 plotvar="var_random_p95", pcat=TRUE)
drawmap(data=plotspatial, map=m, regionvar="CP_seq",
 plotvar="var_total_p95", pcat=TRUE)
```

A.3 Quantile-quantile plot with reference bands

In the following, we specified the code needed to construct a quantile-quantile plot for the quantile residuals with reference bands in a distributional regression model. For this example, we have considered a log-normal response `(y)`.

```
# 1) First, we consider the predicted parameters for each
#  individual observation. These predictions are in the
# '..._predict.res file' of the DR model:

predicted=read.table("C:/Users/.../..._MAIN_mu_REGRESSION_
 y_predict.res",header=T)
 n <- nrow(predicted)

# From these, you will be able to evaluate the estimated
# cumulative distribution function (CDF)
# (e.g. the CDF of the log-normal with  the corresponding
# estimated parameters plugged in) on the observed
# responses. These values will then go into the quantile
# residuals  from  which you can produce the QQ-Plot:

qq <- qnorm(plnorm(predicted$y,
 meanlog=predicted$pmean_param_mu,
 sdlog=predicted$pmean_param_sigma))
```

```
aux <- qqnorm(qq, plot.it=FALSE)

# ---------------------------------------------------------
# 2) Second, we construct the bands, following the next
# steps:
# 2.1) To simulate data from the fitted distribution,
# e.g. a log-normal with estimated mu and sigma
# 2.2) The quantile residuals for these new data would
# then be obtained by plugging in the newly simulated
# data into the CDF estimated from the original data

niter <- 1000
bands <- matrix(0, ncol=niter, nrow=n)
for(i in 1:niter)
{
ytilde <- rlnorm(n, meanlog=predicted$pmean_param_mu,
sdlog=predicted$pmean_param_sigma)
bands[,i] <- sort(qnorm(plnorm(ytilde, meanlog=
predicted$pmean_param_mu,
sdlog=predicted$pmean_param_sigma)))
}
min <- apply(bands, 1, min)
max <- apply(bands, 1, max)
```

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Appendix B

Supplementarial material to Chapter 4: "Extensions to bivariate responses: Copula regression models"

The following shows the most important parts of the code created for Chapter 4 of the thesis.

B.1 Frequentist CGAMLSS code

To estimate the model, the R software GJRM package Marra and Radice (2017a) was used.

```
mu1 <-   HbA1c      ~ s(Glucose, bs = "ps", k=10) + s(Age,
 bs = "ps", k=10) + factor(Gender)
 + s(Bmi, bs = "ps", k=10) + s(Mcv, bs = "ps", k=10)
mu2 <- Fructosamine ~ s(Glucose, bs = "ps", k=10) + s(Age,
 bs = "ps", k=10) + factor(Gender)
 + s(Bmi, bs = "ps", k=10) + s(Albumine, bs = "ps", k=10)
sd1   <- ~ s(Glucose, bs = "ps", k=10) + s(Age, bs = "ps",
 k=10) + factor(Gender)
sd2   <- ~ s(Glucose, bs = "ps", k=10) + s(Age, bs = "ps",
 k=10) + factor(Gender)
theta <- ~ s(Glucose, bs = "ps", k=10) + s(Age,
 bs = "ps", k=10) + factor(Gender) + s(Mcv, bs = "ps",
 k=10)
f  <- list(mu1, mu2, sd1, sd2, theta)
m1 <- gjrm(f, data = data, margins = c("LN", "LN"),
 Model = "B",  BivD="G0")
```

In the GJRM package, the `plot` function represents the centred effects of the response variables. To represent the effects of the continuous covariates at the true scale of the response variables, the function `pred.mvt()` can be used. This function takes into account the link functions contemplated, and allows for the fact that the mean and variability of a distribution may depend on the latter's parameters. (See Marginal Distributions, beginning of Section 4.4.2 of Chapter 4). By way of example, the following shows how to represent the effect of glucose on the mean concentration of HbA1c. The effect of Gender was set to Women (zero), while all continuous covariates but the one being visualised (glucose) were fixed at the average values for the entire data set.

```
glucoses <- seq(min(Glucose), max(Glucose), 1)
nw       <- data.frame(Glucose = glucoses, Gender = 0,
 Age=mean(data$Age), Bmi=mean(data$Bmi),
 Mcv=mean(data$Mcv, na.rm=T))
res <- pred.mvt(m1, eq = 1, fun = "mean", newdata = nw,
 n.sim = 10000, prob.lev = 0.05)
 minimum  <- min(as.numeric(res$CIpred))
 maximum  <- max(as.numeric(res$CIpred))

plot(glucoses, res$pred, type = "l", ylab = "E(HbA1c)",
 xlab = "Glucose (mg/dl)",
 ylim=c(minimum,maximum))
polygon(c(glucoses,rev(glucoses)),c(res$CIpred[, 1],
 rev(res$CIpred[, 2])),col="gray80",border=NA)
lines(glucoses, res$pred, type = "l")
```

In the function `pred.mvt`, "`eq`" can take values of 1 (referring to the first marginal) or 2 (second marginal). The user must also specify the effect to be visualized with the option "`fun`" which can take the value "`mean`", "`variance`" or "`tau`".

B.2 Bayesian CGAMLSS code

BayesX (Belitz et al., 2015) software was used for this process.

```
% Dataset
dataset d
d.infile using /home/jdomin/SMMR/data.raw % The correct
% path must be written here
```

```
d.replace HbA1c        = log(HbA1c)
d.replace Fructosamine = log(Fructosamine)
mcmcreg yreg
yreg.outfile = /home/jdomin/SMMR/model

%Model Estimation
yreg.hregress HbA1c  = const+Gender+
 Glucose(pspline, nrknots=10, lambda=1000)+
 Age(pspline, nrknots=10, lambda=1000),
 copula family=normal equationtype=sigma2
 iterations=12000 step=10 burnin=2000 using d
yreg.hregress HbA1c  = const+Gender+
 Glucose(pspline, nrknots=10, lambda=1000)+
 Age(pspline, nrknots=10, lambda=1000)+
 Bmi(pspline, nrknots=10, lambda=1000)+
 Mcv(pspline,nrknots=10, lambda=1000),
 family=normal equationtype=mu using d

yreg.hregress Fructosamine  = const+Gender+
 Glucose(pspline, nrknots=10, lambda=1000)+
 Age(pspline, nrknots=10, lambda=1000),
 family=normal equationtype=sigma2  using d
yreg.hregress Fructosamine  = const+Gender+
 Glucose(pspline, nrknots=10, lambda=1000)+
 Age(pspline, nrknots=10, lambda=1000) +
 Bmi(pspline, nrknots=10, lambda=1000) +
 Albumine(pspline,nrknots=10, lambda=1000),
 family=normal equationtype=mu using d
yreg.hregress Fructosamine  = const+Gender+
 Glucose(pspline, nrknots=10, lambda=1000) +
 Age(pspline, nrknots=10, lambda=1000) +
 Mcv(pspline, nrknots=10, lambda=1000),
 predict=light family=clayton_copula
 equationtype=rho setseed=123  using d
drop yreg
```

R software can be used to represent the results obtained with BayesX. In R, the `plot2d` functions of the `R2BayesX` package, and the `plotnonp` function of the `BayesX` package, allow the effects of the centred continuous covariates to be represented. To do this at the true scale of the response variables, some code was created by the author of this thesis. By way of example, the following shows how to represent the effect of glucose on the mean concentration of HbA1c:

```
# ------------------------------------------------------------
library("BayesX")
library("splines")
# ------------------------------------------------------------

fixed_mu1_samples <- read.table("model_MAIN_mu_REGRESSION
 _HbA1c_LinearEffects_sample.raw",
 header=TRUE)[,-1,drop=FALSE]

fixed_sigma1_samples <- read.table("model_MAIN_sigma2_
 REGRESSION_HbA1c_LinearEffects_sample.raw", header=TRUE)
 [,-1,drop=FALSE]

f11 <- read.table("model_MAIN_mu_REGRESSION_HbA1c_nonlinear
 _pspline_effect_of_Glucose_sample.raw",
 header=TRUE)[,-1]
f12 <- read.table("model_MAIN_mu_REGRESSION_HbA1c_nonlinear
 _pspline_effect_of_Age_sample.raw",
 header=TRUE)[,-1]
f13 <- read.table("model_MAIN_mu_REGRESSION_HbA1c_nonlinear
 _pspline_effect_of_Bmi_sample.raw",
 header=TRUE)[,-1]
f14 <- read.table("model_MAIN_mu_REGRESSION_HbA1c_nonlinear
 _pspline_effect_of_Mcv_sample.raw",
 header=TRUE)[,-1]

v11 <- read.table("model_MAIN_sigma2_REGRESSION_HbA1c_
 nonlinear_pspline_effect_of_Glucose_sample.raw",
 header=TRUE)[,-1]
v12 <- read.table("model_MAIN_sigma2_REGRESSION_HbA1c_
 nonlinear_pspline_effect_of_Age_sample.raw",
 header=TRUE)[,-1]

fixed_B1 <- cbind(rep(1,100), rep(1,100))
fixed_B0 <- cbind(rep(1,100), rep(0,100))
Glucose_seq <- seq(min(data$Glucose), max(data$Glucose),
 length=100)
source("model_MAIN_mu_REGRESSION_HbA1c_nonlinear_pspline_
 effect_of_Glucose_basisR.res")
Glucose_B1=BayesX.design.matrix(Glucose_seq)

Age_seq <- rep(mean(unique(data$Age)), 100)
source("model_MAIN_mu_REGRESSION_HbA1c_nonlinear_pspline_
```

```
 effect_of_Age_basisR.res")
Age_B1=BayesX.design.matrix(Age_seq)

Bmi_seq <- rep(mean(unique(data$Bmi)), 100)
source("model_MAIN_mu_REGRESSION_HbA1c_nonlinear_pspline_
 effect_of_Bmi_basisR.res")
Bmi_B1=BayesX.design.matrix(Bmi_seq)

Mcv_seq <- rep(mean(unique(data$Mcv)), 100)
source("model_MAIN_mu_REGRESSION_HbA1c_nonlinear_pspline_
 effect_of_Mcv_basisR.res")
Mcv_B1=BayesX.design.matrix(Mcv_seq)

Glucose_seq <- seq(min(data$Glucose), max(data$Glucose),
 length=100)
source("model_MAIN_sigma2_REGRESSION_HbA1c_nonlinear_pspline_
 effect_ of_Glucose_basisR.res")
Glucose_B2=BayesX.design.matrix(Glucose_seq)

Age_seq <- rep(mean(unique(data$Age)), 100)
source("model_MAIN_sigma2_REGRESSION_HbA1c_nonlinear_pspline_
 effect_of_Age_basisR.res")
Age_B2=BayesX.design.matrix(Age_seq)


niter <- 1000
eta_mu_Glucose1 <- matrix(0, nrow=100, ncol=niter)
eta_sigma2_Glucose1 <- matrix(0, nrow=100, ncol=niter)
for(i in 1:niter)
{
 eta_mu_Glucose1[,i] <- fixed_B0 %*% t(fixed_mu1_samples[i,])+
  Glucose_B1 %*% t(f11[i,])  +  Age_B1 %*% t(f12[i,])+
  Bmi_B1 %*% t(f13[i,])+  Mcv_B1 %*% t(f14[i,])
 eta_sigma2_Glucose1[,i] <- fixed_B0 %*%
  t(fixed_sigma1_samples[i,])
  + Glucose_B2 %*% t(v11[i,])
  + Age_B2 %*% t(v12[i,])
}

lnmean <- function(eta_mu, eta_sigma2){
 exp(eta_mu+0.5*exp(eta_sigma2))
    }
lnvar <- function(eta_mu, eta_sigma2){
```

```
 sqrt((exp(exp(eta_sigma2))-1)*
 (exp(2*eta_mu+exp(eta_sigma2))))
    }

mean_Glucose_HbA1c1 <- lnmean(eta_mu_Glucose1,
 eta_sigma2_Glucose1)
std_Glucose_HbA1c1 <- lnvar(eta_mu_Glucose1,
 eta_sigma2_Glucose1)

mean_Glucose1_mean <- apply(mean_Glucose_HbA1c1 , 1,
 mean)
mean_Glucose1_q2p5 <- apply(mean_Glucose_HbA1c1 , 1,
 quantile, prob=0.025)
mean_Glucose1_q97p5 <- apply(mean_Glucose_HbA1c1 , 1,
 quantile, prob=0.975)

std_Glucose1_mean <- apply(std_Glucose_HbA1c1 , 1,
 mean)
std_Glucose1_q2p5 <- apply(std_Glucose_HbA1c1 , 1,
 quantile, prob=0.025)
std_Glucose1_q97p5 <- apply(std_Glucose_HbA1c1 , 1,
 quantile, prob=0.975)

plot(Glucose_seq, mean_Glucose1_mean, type="l",
 xlab="Glucose", ylab="E(HbA1c)")
polygon(c(Glucose_seq,rev(Glucose_seq)),
 c(mean_Glucose1_q2p5,rev(mean_Glucose1_q97p5)),border=NA)
 lines(Glucose_seq, mean_Glucose1_mean, type="l",
  xlab="Glucose", ylab="E(HbA1c)")

plot(Glucose_seq, std_Glucose1_mean, type="l",
  xlab="Glucose", ylab="SD(HbA1c)")
polygon(c(Glucose_seq,rev(Glucose_seq)),
 c(std_Glucose1_q2p5,rev(std_Glucose1_q97p5)),border=NA)
lines(Glucose_seq, std_Glucose1_mean*, type="l",
 xlab="Glucose", ylab="E(HbA1c)")
```

In the latter code, the effect of Gender was set to Women (zero), while all continuous covariates but the one being visualised were fixed at the average values for the entire data set.

Appendix C

Supplementarial material to Chapter 5: "Distributional regression models including functional data"

C.1 Software and code

Chapter 5 proposes the incorporation of functional data covariates in the framework of a DR. For that purpose, open-source BayesX software (Belitz et al., 2015) could be used. The `BayesX` (Kneib et al., 2014) and `R2BayesX` (Umlauf et al., 2015; Belitz et al., 2016) R packages can also be used as graphic interfaces as showed in Appendix A.

The following shows how to model the effect of a continuous variable on the mean and standard deviation of a response variable `y` in a DR regression model.

```
mcmcreg yreg

yreg.hregress y = const + id(userdefined,penmatdata=penmat,
 designmatdata=designmat,centermethod=meanfd),
 iterations=10000 step=10 burnin=2000
 family=normal2 equationtype=sigma  using d

yreg.hregress y = const + id(userdefined,penmatdata=penmat,
 designmatdata=designmat,centermethod=meanfd),
 family=normal2 predict=light equationtype=mu  using d
```

where `d` is the dataset; `penmat` and `designmat` are the penalty and design matrix, respectively. In practice, for *MCMC*, we chose exactly the same design matrices and penalties as in *boosting* approach to make things comparable between *MCMC* and *boosting*.

In the above code, `id` represent the identification of each individual. As commented before, DR approach allows to define for each parameter of the response distribution an additive predictor. This is done by defining for each parameter of the distribution variable response an equation. Arguments `family` and `equationtype` permit to define the equation type desired by the user.

It should be noted that functional terms can be included in all the equation of the DR model -as in the above code-. Furthermore, these terms can be easily mixed with other types of smooths as non-linear, spatial-temporal or random effects, for instance. The different terms have to be separated by "+" signs in the desired equations. By way of example, we present the following code:

```
mcmcreg yreg

yreg.hregress y = const + id(userdefined,penmatdata=penmat,
 designmatdata=designmat,centermethod=meanfd)
 + x1 + x2(pspline),
 iterations=10000 step=10 burnin=2000
 family=normal2 equationtype=sigma  using d

yreg.hregress y = const + id(userdefined,penmatdata=penmat,
 designmatdata=designmat,centermethod=meanfd)
 + x1 + x2(pspline)
 + district(spatial,map=m) + district(random),
 family=normal2 predict=light equationtype=mu  using d
```

where `x1` is a categorical variable, `x2` is a continuous covariate (which is assumed to have a possibly nonlinear effect on the mean and standard deviation of the response distribution `y`) and `m` denotes a map object. In this case, spatial effect of the district is incorporate as an additional covariate to explain the mean of the response variable, and it was split up into a a spatially correlated part, `district(spatial, map=m)`, and an uncorrelated part, `district(random)`.

To plot the functional effects obtained in BayesX, the following code can be used in R:

```
t <- seq(t1, tR, length.out=R) #index of functional covariate
# R: number of measured points
test <- read.table(paste("path of the results","_MAIN_mu_
 REGRESSION_y_nonlinear_userdefined_effect_of_id_param.res",
 sep=""),header=TRUE)
plot(seq(min(t),max(t),length.out=length(test$pmean)),
 test$pmean, col="grey")
```

Appendix D

Supplementarial material to Chapter 6: "Functional regression CGAMLSS"

D.1 Software and code

All computations of Chapter 6 were performed in R (R Core Team, 2017) using the GJRM package (Marra and Radice, 2017b). This section shows the main code snippets that have been used to fit the model and produce the outputs.

```
> eq1  <- HbA1c        ~ s(z = Time, x = Glu, by = Lr, bs = "dt",
xt = list(tf = list(Glu = "QTransform"), basistype = "te")) +
 s(age)
> eq2  <- fructosamine ~ s(z = Time, x = Glu, by = Lr, bs = "dt",
xt = list(tf = list(Glu = "QTransform"), basistype = "te")) +
  s(age)
> eq12 <-              ~  s(age)
> eq21 <-              ~  s(age)
> eqth <-              ~  s(fpg)
> f.l   <- list(eq1, eq2, eq12, eq21, eqth)
> copm1 <- gjrm(f.l, margins = c("LO", "LN"), BivD = "N",
  Model = "B")

> summary(copm1)
COPULA:   Gaussian
MARGIN 1: logistic
MARGIN 2: log-normal

EQUATION 1
Link function for mu.1: identity
Formula: HbA1c ~ s(z = Time, x = Glu, by = Lr, bs = "dt",
```

```
xt = list(tf = list(Glu = "QTransform"),
basistype = "te", k = 4, m = 2)) + s(age)

Parametric coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)    5.431    112.431   0.048    0.961

Smooth components' approximate significance:
edf Ref.df Chi.sq  p-value
te(Time,Glu):Lr 11.978 12.001  34.58 0.000561 ***
s(age)           8.822  8.988 173.92  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


EQUATION 2
Link function for mu.2: log
Formula: fructosamine ~ s(z = Time, x = Glu, by = Lr, bs = "dt",
xt = list(tf = list(Glu = "QTransform"),
basistype = "te", k = 5, m = 2)) + s(age)

Parametric coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)    5.487     59.926   0.092    0.927

Smooth components' approximate significance:
edf Ref.df Chi.sq  p-value
te(Time,Glu):Lr 20.000 20.002  46.61 0.000665 ***
s(age)           5.629  6.761  14.05 0.040636 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


[...]

EQUATION 5
Link function for theta: atanh
Formula: ~s(fpg)

Parametric coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.01916    0.08895   0.215    0.829

Smooth components' approximate significance:
edf Ref.df Chi.sq p-value
s(glucose)   9      9  24.76 0.00325 **
```

```
[...]
```

The perspective and contour plots shown in Figures 6.4 and 6.5 of Chapter 6 were produced using something like:

```
vis.gjrm(copm1, eq = 1, fun = "mean", view = c("Time",
 "Glu"),
plot.type = "persp", n.grid = 100,  cond = list(age = 46),
 color = "heat",  main = "", xlab = "Time (hours)", ylab =
 "Glucose (mg/dl)", border = "NA", zlab = "E(HbA1c)")
```

where `eq` can be either 1 or 2, and the possible values for `fun` are `mean` and `variance`.

As for the effects of the continuous covariates, the impact of age on the HbA1c mean levels can be produced directly using the standard plotting command available in the package (this is because for a logistic distribution the link function between the mean and its additive predictor is identity) whereas the other effects are produced using the `pred.mvt()` function illustrated below which takes into account the presence of links functions, and that the mean and variance of a distribution can depend on its parameters in a complicated manner (see Table 4.1, Chapter 4).

```
ages <- seq(18, 87, 1)
nw   <- data.frame(age = ages, Glu = 1, Time = mean(Time),
 L = mean(L))
res  <- pred.mvt(copm1, eq = 1, fun = "variance",
 newdata = nw,
 n.sim = 10000,  prob.lev = 0.05)
mi <- min(as.numeric(res$CIpred))
ma <- max(as.numeric(res$CIpred))
plot(ages, res$pred, type = "l", ylim = c(mi, ma), ylab =
 "V(HbA1c)",
xlab = "Age (years)")
polygon(c(ages, rev(ages)), c(res$CIpred[, 1],
 rev(res$CIpred[, 2])),
col = "gray80", border = NA)
lines(ages, res$pred, type = "l", ylim = c(mi, ma), ylab =
 "V(HbA1c)", xlab = "Age")
```

Here the possible values for `fun` are `mean`, `variance` and `tau`.